

Kernel Methods on Riemannian Manifolds with Gaussian RBF Kernels

Sadeep Jayasumana, *Student Member, IEEE*, Richard Hartley, *Fellow, IEEE*,
Mathieu Salzmann, *Member, IEEE*, Hongdong Li, *Member, IEEE*, and
Mehrtash Harandi, *Member, IEEE*

Abstract—In this paper, we develop an approach to exploiting kernel methods with manifold-valued data. In many computer vision problems, the data can be naturally represented as points on a Riemannian manifold. Due to the non-Euclidean geometry of Riemannian manifolds, usual Euclidean computer vision and machine learning algorithms yield inferior results on such data. In this paper, we define Gaussian radial basis function (RBF)-based positive definite kernels on manifolds that permit us to embed a given manifold with a corresponding metric in a high dimensional reproducing kernel Hilbert space. These kernels make it possible to utilize algorithms developed for linear spaces on nonlinear manifold-valued data. Since the Gaussian RBF defined with any given metric is not always positive definite, we present a unified framework for analyzing the positive definiteness of the Gaussian RBF on a generic metric space. We then use the proposed framework to identify positive definite kernels on two specific manifolds commonly encountered in computer vision: the Riemannian manifold of symmetric positive definite matrices and the Grassmann manifold, i.e., the Riemannian manifold of linear subspaces of a Euclidean space. We show that many popular algorithms designed for Euclidean spaces, such as support vector machines, discriminant analysis and principal component analysis can be generalized to Riemannian manifolds with the help of such positive definite Gaussian kernels.

Index Terms—Riemannian manifolds, Gaussian RBF kernels, kernel methods, positive definite kernels, symmetric positive definite matrices, Grassmann manifolds

1 INTRODUCTION

MATHEMATICAL entities that do not form Euclidean spaces but lie on nonlinear manifolds are often encountered in computer vision. Examples include 3D rotation matrices that form the Lie group $SO(3)$, normalized histograms that form the unit n -sphere S^n , symmetric positive definite (SPD) matrices and linear subspaces of a Euclidean space. Recently, the latter two manifolds have drawn significant attention in the computer vision community due to their widespread applications. SPD matrices, which form a Riemannian manifold when endowed with an appropriate metric [1], are encountered in computer vision in the forms of covariance region descriptors [2], [3], diffusion tensors [1], [4] and structure tensors [5], [6]. Linear subspaces of a Euclidean space, known to form a Riemannian manifold named the Grassmann manifold, are commonly used to model image sets [7], [8] and videos [9].

Since manifolds lack a vector space structure and other Euclidean structures such as norm and inner product, many popular computer vision and machine learning algorithms including support vector machines (SVM), principal component analysis (PCA) and mean-shift clustering cannot be

applied in their original forms on manifolds. One way of dealing with this difficulty is to neglect the nonlinear geometry of manifold-valued data and apply Euclidean methods directly. As intuition suggests, this approach often yields poor accuracy and undesirable effects, see, e.g., [1], [10] for SPD matrices.

When the manifold under consideration is Riemannian, another common approach used to cope with its nonlinearity consists in approximating the manifold-valued data with its projection to a tangent space at a particular point on the manifold, for example, the mean of the data. Such tangent space approximations are often calculated successively as the algorithm proceeds [3]. However, mapping data to a tangent space only yields a first-order approximation of the data that can be distorted, especially in regions far from the origin of the tangent space. Moreover, iteratively mapping back and forth to the tangent spaces significantly increases the computational cost of the algorithm. It is also difficult to choose the origin of the tangent space, which heavily affects the accuracy of this approximation.

In Euclidean spaces, the success of many computer vision algorithms arises from the use of kernel methods [11], [12]. Therefore, one could think of following the idea of kernel methods in \mathbb{R}^n and embed a manifold in a high dimensional Reproducing Kernel Hilbert Space (RKHS), where linear geometry applies. Many Euclidean algorithms can be directly generalized to an RKHS, which is a vector space that possesses an important structure: the inner product. Such an embedding, however, requires a kernel function defined on the manifold, which, according to Mercer's theorem [12], should be positive definite. While

• The authors are with the College of Engineering and Computer Science, Australian National University, Canberra and NICTA, Canberra.
E-mail: {sadeep.jayasumana, richard.hartley, Hongdong.Li}@anu.edu.au, {mathieu.salzmann, mehrtash.harandi}@nicta.com.au.

Manuscript received 14 Nov. 2013; revised 6 Feb. 2015; accepted 9 Mar. 2015.
Date of publication 17 Mar. 2015; date of current version 6 Nov. 2015.

Recommended for acceptance by M. Belkin.

For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.

Digital Object Identifier no. 10.1109/TPAMI.2015.2414422

many positive definite kernels are known for Euclidean spaces, such knowledge remains limited for manifold-valued data.

The Gaussian radial basis function (RBF) kernel $\exp(-\gamma \|x - y\|^2)$ is perhaps the most popular and versatile kernel in Euclidean spaces. It would therefore seem attractive to generalize this kernel to manifolds by replacing the Euclidean distance in the RBF by a more accurate nonlinear distance measure on the manifold, such as the geodesic distance. However, a kernel formed in this manner is *not* positive definite in general.

In this paper, we aim to generalize the successful and powerful kernel methods to manifold-valued data. To this end, we analyze the Gaussian RBF kernel on a generic manifold and provide necessary and sufficient conditions for the Gaussian RBF kernel generated by a distance function on any nonlinear manifold to be positive definite. This lets us generalize kernel methods to manifold-valued data while preserving the favorable properties of the original algorithms.

We apply our theory to analyze the positive definiteness of Gaussian kernels defined on two specific manifolds: the Riemannian manifold of SPD matrices and the Grassmann manifold. Given the resulting positive definite kernels, we discuss different kernel methods on these two manifolds including, kernel SVM, multiple kernel learning (MKL) and kernel PCA. Our experiments on a variety of computer vision tasks, such as pedestrian detection, segmentation, face recognition and action recognition, show that our manifold kernel methods outperform the corresponding Euclidean algorithms that neglect the manifold geometry, as well as other state-of-the-art techniques specifically designed for manifolds.

2 RELATED WORK

In this paper, we focus on the Riemannian manifold of SPD matrices and on the Grassmann manifold. SPD matrices find a variety of applications in computer vision [13]. For instance, covariance region descriptors are used in object detection [3], texture classification [2], [14], object tracking, action recognition and human recognition [15], [16]. Diffusion Tensor Imaging (DTI) was one of the pioneering fields for the development of non-linear algorithms on SPD matrices [1], [10]. In optical flow estimation and motion segmentation, structure tensors are often employed to encode important image features, such as texture and motion [5], [6]. Structure tensors have also been used in single image segmentation [17].

Grassmann manifolds are widely used to encode image sets and videos for face recognition [7], [8], activity recognition [9], [18], and motion grouping [18]. In image set based face recognition, a set of face images of the same person is represented as a linear subspace, hence a point on a Grassmann manifold. In activity recognition and motion grouping, a subspace is formed either directly from the sequence of images containing a specific action, or from the parameters of a dynamic model obtained from the sequence [9].

In recent years, several optimization algorithms have been proposed for Riemannian manifolds. In particular, LogitBoost for classification on Riemannian manifolds was introduced in [3]. This algorithm has the drawbacks of

approximating the manifold by tangent spaces and not scaling with the number of training samples due to the heavy use of exponential/logarithmic maps to transit between the manifold and the tangent space, as well as of gradient descent based Karcher mean calculation. Here, our positive definite kernels enable us to use more efficient and accurate classification algorithms on manifolds without requiring tangent space approximations. Furthermore, as shown in [19], [20], extending existing kernel-free, manifold-based binary classifiers to the multi-class case is not straightforward. In contrast, the kernel-based classifiers on manifolds described in this paper can readily be used in multi-class scenarios.

In [5], dimensionality reduction and clustering methods were extended to manifolds by designing Riemannian versions of Laplacian Eigenmaps (LE), Locally Linear Embedding (LLE) and Hessian LLE (HLLE). Clustering was performed after mapping to a low dimensional space which does not necessarily preserve all the information in the original data. Instead, we use our kernels to perform clustering in a higher dimensional RKHS that embeds the manifold of interest.

The use of kernels on SPD matrices has previously been advocated for locality preserving projections [21] and sparse coding [15]. In the first case, the kernel, derived from the affine-invariant distance, is not positive definite in general [21]. In the second case, the kernel is positive definite only for some values of the Gaussian bandwidth parameter γ [15]. For all kernel methods, the optimal choice of γ largely depends on the data distribution and hence constraints on γ are not desirable. Moreover, many popular automatic model selection methods require γ to be continuously variable [22].

In [7], [23], the Projection kernel and its extensions were introduced and employed for classification on Grassmann manifolds. While those kernels are analogous to the linear kernel in Euclidean spaces, our kernels are analogous to the Gaussian RBF kernel. In Euclidean spaces, the Gaussian kernel has proven more powerful and versatile than the linear kernel. As shown in our experiments, this also holds for kernels on manifolds.

Recently, mean-shift clustering with the heat kernel on Riemannian manifolds was introduced [24]. However, due to the mathematical complexity of the kernel function, computing the exact kernel is not tractable and hence only an approximation of the true kernel was used. Parallel to our work, kernels on SPD matrices and on Grassmann manifolds were used in [25], albeit without explicit proof of their positive definiteness. In contrast, in this paper, we introduce a unified framework for analyzing the positive definiteness of the Gaussian kernel defined on any manifold and use this framework to identify provably positive definite kernels on the manifold of SPD matrices and on the Grassmann manifold.

Other than for satisfying Mercer's theorem to generate a valid RKHS, positive definiteness of the kernel is a required condition for the convergence of many kernel based algorithms. For instance, the support vector machine learning problem is convex only when the kernel is positive definite [26]. Similarly, positive definiteness of all participating kernels is required to guarantee the convexity in multiple kernel learning [27]. Although theories have been proposed to exploit non-positive definite kernels [28], [29], they have not experienced widespread success. Many of these methods

first enforce positive definiteness of the kernel matrix by flipping or shifting its negative eigenvalues [29]. As a consequence, they result in a distortion of information and become inapplicable with large size kernels, which are not uncommon in learning problems.

It is important to note the difference between this work and manifold-learning methods such as [30]. We work with data sampled from a manifold whose geometry is well known. In contrast, manifold-learning methods attempt to learn the structure of an underlying unknown manifold from data samples. Furthermore, those methods often assume that noise-free data samples lie on a manifold from which noise push them away. In our study, data points, regardless of their noise content, always lie on the mathematically well-defined manifold.

3 MANIFOLDS IN COMPUTER VISION

A *topological manifold*, generally referred to as simply a *manifold*, is a topological space that is locally homeomorphic to the n -dimensional Euclidean space \mathbb{R}^n , for some n . Here, n is referred to as the dimensionality of the manifold. A *differentiable manifold* is a topological manifold that has a globally defined differential structure. The tangent space at a given point on a differentiable manifold is a vector space that consists of the tangent vectors of all possible curves passing through the point.

A *Riemannian manifold* is a differentiable manifold equipped with a smoothly varying inner product on each tangent space. The family of inner products on all tangent spaces is known as the *Riemannian metric* of the manifold. It enables us to define various geometric notions on the manifold such as the angle between two curves and the length of a curve. The *geodesic distance* between two points on the manifold is defined as the length of the shortest curve connecting the two points. Such shortest curves are known as *geodesics* and are analogous to straight lines in \mathbb{R}^n .

The geodesic distance induced by the Riemannian metric is the most natural measure of dissimilarity between two points lying on a Riemannian manifold. However, in practice, many other nonlinear distances or *metrics* which do not necessarily arise from Riemannian metrics can also be useful for measuring dissimilarity on manifolds. It is worth noting that the term *Riemannian metric* refers to a family of inner products while the term *metric* refers to a distance function that satisfies the four metric axioms. A nonempty set endowed with a metric is known as a *metric space* which is a more abstract space than a Riemannian manifold.

In the following, we discuss two important Riemannian manifolds commonly found in computer vision.

3.1 The Riemannian Manifold of SPD Matrices

A $d \times d$, real symmetric positive definite matrix S has the property: $\mathbf{x}^T S \mathbf{x} > 0$ for all nonzero $\mathbf{x} \in \mathbb{R}^d$. The space of $d \times d$ SPD matrices, which we denote by Sym_d^+ , is clearly not a vector space since an SPD matrix when multiplied by a negative scalar is no longer SPD. Instead, Sym_d^+ forms a convex cone in the d^2 -dimensional Euclidean space.

The geometry of the space Sym_d^+ is best explained with a Riemannian metric which induces an infinite distance between an SPD matrix and a non-SPD matrix [1], [31].

Therefore, the geodesic distance induced by such a Riemannian metric is a more accurate distance measure on Sym_d^+ than the Euclidean distance in the d^2 -dimensional Euclidean space it is embedded in. Two popular Riemannian metrics proposed on Sym_d^+ are the affine-invariant Riemannian metric [1] and the log-Euclidean Riemannian metric [10], [31]. They result in the affine-invariant geodesic distance and the log-Euclidean geodesic distance, respectively. These two distances are so far the most widely used metrics on Sym_d^+ .

Apart from these two geodesic distances, a number of other metrics have been proposed for Sym_d^+ to capture its nonlinearity [32]. For a detailed review of metrics on Sym_d^+ , we refer the reader to [33].

3.2 The Grassmann Manifold

A point on the (n, r) Grassmann manifold, where $n > r$, is an r -dimensional linear subspace of the n -dimensional Euclidean space. Such a point is generally represented by an $n \times r$ matrix Y whose columns store an orthonormal basis of the subspace. With this representation, the point on the Grassmann manifold is the subspace spanned by the columns of Y or $\text{span}(Y)$ which we denote by $[Y]$. We use \mathcal{G}_n^r to denote the (n, r) Grassmann manifold.

The set of $n \times r$ ($n > r$) matrices with orthonormal columns forms a manifold known as the (n, r) *Stiefel manifold*. From its embedding in the nr -dimensional Euclidean space, the (n, r) Stiefel manifold inherits a canonical Riemannian metric [34]. Points on \mathcal{G}_n^r are equivalence classes of $n \times r$ matrices with orthonormal columns, where two matrices are equivalent if their columns span the same r -dimensional subspace. Thus, the orthogonal group \mathcal{O}_r acts via isometries (change of orthogonal basis) on the Stiefel manifold by multiplication on the right, and \mathcal{G}_n^r can be identified as the set of orbits of this action. Since this action is both free and proper, \mathcal{G}_n^r forms a manifold, and it is given a Riemannian structure by equipping it with the standard *normal* Riemannian metric derived from the metric on the Stiefel manifold.

A geodesic distance on the Grassmann manifold, called the arc length distance, can be derived from its canonical geometry described above. The arc length distance between two points on the Grassmann manifold turns out to be the l_2 norm of the vector formed by the principal angles between the two subspaces. Several other metrics on this manifold can be derived from the principal angles. We refer the reader to [34] for more details and properties of different Grassmann metrics.

4 HILBERT SPACE EMBEDDING OF MANIFOLDS

An *inner product space* is a vector space equipped with an inner product. A *Hilbert space* is an (often infinite-dimensional) inner product space which is complete with respect to the norm induced by the inner product. A Reproducing Kernel Hilbert Space is a special kind of Hilbert space of functions on some nonempty set \mathcal{X} in which all evaluation functionals are bounded and hence continuous [35]. The inner product of an RKHS of functions on \mathcal{X} can be defined by a bivariate function on $\mathcal{X} \times \mathcal{X}$, known as the *reproducing kernel* of the RKHS.

Many useful computer vision and machine learning algorithms developed for Euclidean spaces depend only

on the notion of inner product, which allows us to measure angles and also distances. Therefore, such algorithms can be extended to Hilbert spaces without effort. A notable special case arises with RKHSs where the inner product of the Hilbert space can be evaluated using a kernel function without computing the actual vectors. This concept, known as the *kernel trick*, is commonly utilized in machine learning in the following setting: input data in some n -dimensional Euclidean space \mathbb{R}^n are mapped to a high dimensional RKHS where some learning algorithm, which requires only the inner product, is applied. We never need to calculate actual vectors in the RKHS since the learning algorithm only requires the inner product of the RKHS, which can be calculated by means of a kernel function defined on $\mathbb{R}^n \times \mathbb{R}^n$. A variety of algorithms can be used with the kernel trick, such as support vector machines, principal component analysis, Fisher discriminant analysis (FDA), k -means clustering and ridge regression.

Embedding lower dimensional data in a higher dimensional RKHS is commonly employed with data that lies in a Euclidean space. The theoretical concepts of such embeddings can directly be extended to manifolds. Points on a manifold \mathcal{M} are mapped to elements in a high (possibly infinite) dimensional Hilbert space \mathcal{H} , a subspace of the space spanned by real-valued functions¹ on \mathcal{M} . A kernel function $k : (\mathcal{M} \times \mathcal{M}) \rightarrow \mathbb{R}$ is used to define the inner product on \mathcal{H} , thus making it an RKHS. The technical difficulty in utilizing Hilbert space embeddings with manifold-valued data arises from the fact that, according to Mercer's theorem, the kernel must be positive definite to define a valid RKHS. While many positive definite kernel functions are known for \mathbb{R}^n , generalizing them to manifolds is not straightforward.

Identifying such positive definite kernel functions on manifolds would, however, be greatly beneficial. Indeed, embedding a manifold in an RKHS has two major advantages: First, the mapping transforms the nonlinear manifold into a (linear) Hilbert space, thus making it possible to utilize algorithms designed for linear spaces with manifold-valued data. Second, as evidenced by the theory of kernel methods in Euclidean spaces, it yields a much richer high-dimensional representation of the original data, making tasks such as classification easier.

In the following we build a framework that enables us to define positive definite kernels on manifolds.

5 THEORY OF POSITIVE AND NEGATIVE DEFINITE KERNELS

In this section, we present some general results on positive and negative definite kernels. These results will be useful for our derivations in later sections. We start with the definition of real-valued positive and negative definite kernels on a set [36]. Note that by the term *kernel* we mean a real-valued bivariate function hereafter.

1. We limit the discussion to real Hilbert spaces and real-valued kernels, since they are the most useful kind in learning algorithms. However, the theory holds for complex Hilbert spaces and complex-valued kernels as well.

Definition 5.1. Let \mathcal{X} be a nonempty set. A kernel $f : (\mathcal{X} \times \mathcal{X}) \rightarrow \mathbb{R}$ is called *positive definite* if it is symmetric (i.e., $f(x, y) = f(y, x)$ for all $x, y \in \mathcal{X}$) and

$$\sum_{i,j=1}^m c_i c_j f(x_i, x_j) \geq 0$$

for all $m \in \mathbb{N}$, $\{x_1, \dots, x_m\} \subseteq \mathcal{X}$ and $\{c_1, \dots, c_m\} \subseteq \mathbb{R}$. The kernel f is called *negative definite* if it is symmetric and

$$\sum_{i,j=1}^m c_i c_j f(x_i, x_j) \leq 0$$

for all $m \in \mathbb{N}$, $\{x_1, \dots, x_m\} \subseteq \mathcal{X}$ and $\{c_1, \dots, c_m\} \subseteq \mathbb{R}$ with $\sum_{i=1}^m c_i = 0$.

It is important to note the additional constraint on $\sum c_i$ for the negative definite case. Due to this constraint, some authors refer to this latter kind as *conditionally negative definite*. However, in this paper, we stick to the most common terminology used in the literature.

We next present the following theorem which plays a central role in this paper. It was introduced by Schoenberg in 1938 [37], well before the theory of Reproducing Kernel Hilbert Spaces was established in 1950 [35].

Theorem 5.2. Let \mathcal{X} be a nonempty set and $f : (\mathcal{X} \times \mathcal{X}) \rightarrow \mathbb{R}$ be a kernel. The kernel $\exp(-\gamma f(x, y))$ is positive definite for all $\gamma > 0$ if and only if f is negative definite.

Proof. We refer the reader to Theorem 3.2.2 of [36] for a detailed proof of this theorem. \square

Note that this theorem describes Gaussian RBF-like exponential kernels that are positive definite for all $\gamma > 0$. One might also be interested in exponential kernels that are positive definite for only some values of γ . Such kernels, for instance, were exploited in [15]. To our knowledge, there is no general result characterizing this kind of kernels. However, the following result can be obtained from the above theorem.

Theorem 5.3. Let \mathcal{X} be a nonempty set and $f : (\mathcal{X} \times \mathcal{X}) \rightarrow \mathbb{R}$ be a kernel. If the kernel $\exp(-\gamma f(x, y))$ is positive definite for all $\gamma \in (0, \delta)$ for some $\delta > 0$, then it is positive definite for all $\gamma > 0$.

Proof. If $\exp(-\gamma f(x, y))$ is positive definite for all $\gamma \in (0, \delta)$, it directly follows from Definition 5.1 that $1 - \exp(-\gamma f(x, y))$ is negative definite for all $\gamma \in (0, \delta)$. Therefore, the pointwise limit

$$\lim_{\gamma \rightarrow 0^+} \frac{1 - \exp(-\gamma f(x, y))}{\gamma} = f(x, y)$$

is also negative definite. Now, since $f(x, y)$ is negative definite, it follows from Theorem 5.2 that $\exp(-\gamma f(x, y))$ is positive definite for all $\gamma > 0$. \square

Next, we highlight an interesting property of negative definite kernels. It is well known that a positive definite kernel represents the inner product of an RKHS [35]. Similarly, a negative definite kernel represents the squared norm of a Hilbert space under some conditions stated by the following theorem.

Theorem 5.4. Let \mathcal{X} be a nonempty set and $f(x, y) : (\mathcal{X} \times \mathcal{X}) \rightarrow \mathbb{R}$ be a negative definite kernel. Then, there exists a Hilbert space \mathcal{H} and a mapping $\psi : \mathcal{X} \rightarrow \mathcal{H}$ such that,

$$f(x, y) = \|\psi(x) - \psi(y)\|^2 + h(x) + h(y)$$

where $h : \mathcal{X} \rightarrow \mathbb{R}$ is a function which is nonnegative whenever f is. Furthermore, if $f(x, x) = 0$ for all $x \in \mathcal{X}$, then $h = 0$.

Proof. The proof for a more general version of this theorem can be found in Proposition 3.3.2 of [36]. \square

Now, we state and prove a lemma that will be useful for the proof of our main theorem.

Lemma 5.5. Let \mathcal{X} be a nonempty set, \mathcal{V} be an inner product space, and $\psi : \mathcal{X} \rightarrow \mathcal{V}$ be a function. Then, $f : (\mathcal{X} \times \mathcal{X}) \rightarrow \mathbb{R}$ defined by $f(x, y) := \|\psi(x) - \psi(y)\|_{\mathcal{V}}^2$ is negative definite.

Proof. The kernel f is obviously symmetric. Based on Definition 5.1, we then need to prove that $\sum_{i,j=1}^m c_i c_j f(x_i, x_j) \leq 0$ for all $m \in \mathbb{N}$, $\{x_1, \dots, x_m\} \subseteq \mathcal{X}$ and $\{c_1, \dots, c_m\} \subseteq \mathbb{R}$ with $\sum_{i=1}^m c_i = 0$. Now,

$$\begin{aligned} \sum_{i,j=1}^m c_i c_j f(x_i, x_j) &= \sum_{i,j=1}^m c_i c_j \|\psi(x_i) - \psi(x_j)\|_{\mathcal{V}}^2 \\ &= \sum_{i,j=1}^m c_i c_j \langle \psi(x_i) - \psi(x_j), \psi(x_i) - \psi(x_j) \rangle_{\mathcal{V}} \\ &= \sum_{j=1}^m c_j \sum_{i=1}^m c_i \langle \psi(x_i), \psi(x_i) \rangle_{\mathcal{V}} \\ &\quad - 2 \sum_{i,j=1}^m c_i c_j \langle \psi(x_i), \psi(x_j) \rangle_{\mathcal{V}} \\ &\quad + \sum_{i=1}^m c_i \sum_{j=1}^m c_j \langle \psi(x_j), \psi(x_j) \rangle_{\mathcal{V}} \\ &= -2 \sum_{i,j=1}^m c_i c_j \langle \psi(x_i), \psi(x_j) \rangle_{\mathcal{V}} \\ &= -2 \left\| \sum_{i=1}^m c_i \psi(x_i) \right\|_{\mathcal{V}}^2 \leq 0. \end{aligned}$$

\square

5.1 Test for the Negative Definiteness of a Kernel

In linear algebra, an $m \times m$ real symmetric matrix M is called *positive semi-definite* if $\mathbf{c}^T M \mathbf{c} \geq 0$ for all $\mathbf{c} \in \mathbb{R}^m$ and *negative semi-definite* if $\mathbf{c}^T M \mathbf{c} \leq 0$ for all $\mathbf{c} \in \mathbb{R}^m$. These conditions are respectively equivalent to M having non-negative eigenvalues and non-positive eigenvalues. Furthermore, a matrix M which has the property $\hat{\mathbf{c}}^T M \hat{\mathbf{c}} \leq 0$ for all $\hat{\mathbf{c}} \in \mathbb{R}^m$ with $\sum \hat{c}_i = 0$, where \hat{c}_i s are the components of the vector $\hat{\mathbf{c}}$, is termed *conditionally negative semi-definite*.

Positive (resp. negative) definiteness of a given kernel—a bivariate function—is usually tested by evaluating the positive semi-definiteness (resp. conditionally negative semi-definiteness) of kernel matrices generated with the kernel.²

2. There is an unfortunate confusion of terminology here. A matrix generated by a positive definite kernel is positive semi-definite while a matrix generated by a negative definite kernel is conditionally negative semi-definite.

Although such a test is not always conclusive if it passes, it is particularly useful to identify non-positive definite and non-negative definite kernels. Given a set of points $\{x_i\}_{i=1}^m \subseteq \mathcal{X}$ and a kernel $k : (\mathcal{X} \times \mathcal{X}) \rightarrow \mathbb{R}$, the kernel matrix K of the given points has entries $K_{ij} = k(x_i, x_j)$. Any kernel matrix generated by a positive (resp. negative) definite kernel must be positive semi-definite (resp. conditionally negative semi-definite). As noted above, it is straightforward to check the positive semi-definiteness of a kernel matrix K by checking its eigenvalues. However, due to the additional constraint that $\sum \hat{c}_i = 0$, checking the conditionally negative semi-definiteness is not straightforward. We therefore suggest the following procedure.

Let $P = I_m - \frac{1}{m} \mathbf{1}_m \mathbf{1}_m^T$, where I_m is the $m \times m$ identity matrix and $\mathbf{1}_m$ is the m -vector of ones. Any given $\hat{\mathbf{c}} \in \mathbb{R}^m$ with $\sum \hat{c}_i = 0$ can be written as $\hat{\mathbf{c}} = P\mathbf{c}$ for some $\mathbf{c} \in \mathbb{R}^m$. Therefore, the condition $\hat{\mathbf{c}}^T M \hat{\mathbf{c}} \leq 0$ is equivalent to $\mathbf{c}^T P M P \mathbf{c} \leq 0$. Hence, we conclude that an $m \times m$ matrix M is conditionally negative semi-definite if and only if $P M P$ is negative semi-definite (i.e., has non-positive eigenvalues). This gives a convenient test for the conditionally negative semi-definiteness of a matrix, which in turn is useful to evaluate the negative definiteness of a given kernel.

6 KERNELS ON MANIFOLDS

A number of well-known kernels exist for \mathbb{R}^n including the linear kernel, polynomial kernels and the Gaussian RBF kernel. The key challenge in generalizing kernel methods from Euclidean spaces to manifolds lies in defining appropriate positive definite kernels on the manifold. There is no straightforward way to generalize Euclidean kernels such as the linear kernel and polynomial kernels to nonlinear manifolds, since these kernels depend on the linear geometry of \mathbb{R}^n . However, we show that the popular Gaussian RBF kernel can be generalized to manifolds under certain conditions.

In this section, we first introduce a general theorem that provides necessary and sufficient conditions to define a positive definite Gaussian RBF kernel on a given manifold and then show that some popular metrics on Sym_d^+ and \mathcal{G}_n^r yield positive definite Gaussian RBFs on the respective manifolds.

6.1 The Gaussian RBF Kernel on Metric Spaces

The Gaussian RBF kernel has proven very effective in Euclidean spaces for a variety of kernel-based algorithms. It maps the data points to an infinite dimensional Hilbert space, which, intuitively, yields a very rich representation of the data. In \mathbb{R}^n , the Gaussian kernel can be expressed as $k_G(\mathbf{x}, \mathbf{y}) := \exp(-\gamma \|\mathbf{x} - \mathbf{y}\|^2)$, which makes use of the Euclidean distance between two data points \mathbf{x} and \mathbf{y} . To define a kernel on a manifold, we would like to replace the Euclidean distance by a more accurate distance measure on the manifold. However, not all geodesic distances yield positive definite kernels. For example, in the case of the unit n -sphere embedded in \mathbb{R}^{n+1} , $\exp(-\gamma d_g^2(x, y))$, where d_g is the usual geodesic distance on the manifold, is not positive definite.

We now state our main theorem which provides the necessary and sufficient conditions to obtain a positive definite

Gaussian kernel from a given distance function defined on a generic space.

Theorem 6.1. *Let (M, d) be a metric space and define $k : (M \times M) \rightarrow \mathbb{R}$ by $k(x, y) := \exp(-\gamma d^2(x, y))$. Then, k is a positive definite kernel for all $\gamma > 0$ if and only if there exists an inner product space \mathcal{V} and a function $\psi : M \rightarrow \mathcal{V}$ such that $d(x, y) = \|\psi(x) - \psi(y)\|_{\mathcal{V}}$.*

Proof. We first note that positive definiteness of $k(\cdot, \cdot)$ for all γ and negative definiteness of $d^2(\cdot, \cdot)$ are equivalent conditions according to Theorem 5.2.

To prove the forward direction of the present theorem, let us first assume that \mathcal{V} and ψ exist such that $d(x, y) = \|\psi(x) - \psi(y)\|_{\mathcal{V}}$. Then, from Lemma 5.5, d^2 is negative definite and hence k is positive definite for all γ .

On the other hand, if k is positive definite for all γ , then d^2 is negative definite. Furthermore, $d(x, x) = 0$ for all $x \in M$ since d is a metric. Following Theorem 5.4, then \mathcal{V} and ψ exist such that $d(x, y) = \|\psi(x) - \psi(y)\|_{\mathcal{V}}$. \square

6.2 Geodesic Distances and the Gaussian RBF

A Riemannian manifold, when considered with its geodesic distance, forms a metric space. Given Theorem 6.1, it is then natural to wonder under which conditions would a geodesic distance on a manifold yield a Gaussian RBF kernel. We now present and prove the following theorem, which answers this question for complete Riemannian manifolds.

Theorem 6.2. *Let \mathcal{M} be a complete Riemannian manifold and d_g be the geodesic distance induced by its Riemannian metric. The Gaussian RBF kernel $k_g : (\mathcal{M} \times \mathcal{M}) \rightarrow \mathbb{R} : k_g(x, y) := \exp(-\gamma d_g^2(x, y))$ is positive definite for all $\gamma > 0$ if and only if \mathcal{M} is isometric (in the Riemannian sense) to some Euclidean space \mathbb{R}^n .*

Proof. If \mathcal{M} is isometric to some \mathbb{R}^n , the geodesic distance on \mathcal{M} is simply the Euclidean distance in \mathbb{R}^n , which can be trivially shown to yield a positive definite Gaussian RBF kernel by setting ψ in Theorem 6.1 to the identity function.

On the other hand, if k_g is positive definite, from Theorem 6.1, there exists an inner product space \mathcal{V}_g and a function $\psi_g : \mathcal{M} \rightarrow \mathcal{V}_g$ such that $d_g(x, y) = \|\psi_g(x) - \psi_g(y)\|_{\mathcal{V}_g}$. Let \mathcal{H}_g be the completion of \mathcal{V}_g . Therefore, \mathcal{H}_g is a Hilbert space, in which \mathcal{V}_g is dense.

Now, take any two points x_0, x_1 in \mathcal{M} . Since the manifold is complete, from the Hopf-Rinow theorem, there exists a geodesic $\delta(t)$ joining them with $\delta(0) = x_0$ and $\delta(1) = x_1$, and realizing the geodesic distance. By definition, $\delta(t)$ has a constant speed $d_g(x_0, x_1)$. Therefore, for all $x_t = \delta(t)$ where $t \in [0, 1]$, the following equality holds

$$d_g(x_0, x_t) + d_g(x_t, x_1) = d_g(x_0, x_1).$$

This must also be true for images $\psi(x_t)$ in \mathcal{H}_g for $t \in [0, 1]$. However, since \mathcal{H}_g is a Hilbert space, this is only possible if all the points $\psi(x_t)$ lie on a straight line in \mathcal{H}_g . Let $\psi(\mathcal{M})$ be the range of ψ . From the previous argument, for any two points in $\psi(\mathcal{M}) \subseteq \mathcal{H}_g$, the straight line segment joining them is also in $\psi(\mathcal{M})$. Therefore, $\psi(\mathcal{M})$

is a convex set in \mathcal{H}_g . Now, since \mathcal{M} is complete, any geodesic must be extensible indefinitely. Therefore, the corresponding line segment in $\psi(\mathcal{M})$ must also be extensible indefinitely. This proves that $\psi(\mathcal{M})$ is an affine subspace of \mathcal{H}_g , which is isometric to \mathbb{R}^n , for some n . Since \mathcal{M} is isometric to $\psi(\mathcal{M})$, this proves that \mathcal{M} is isometric to the Euclidean space \mathbb{R}^n . \square

According to Theorem 6.2, it is possible to obtain a positive definite Gaussian kernel from the geodesic distance on a Riemannian manifold only when the manifold is made essentially equivalent to some \mathbb{R}^n by the Riemannian metric that defines the geodesic distance. Although this is possible for some Riemannian manifolds, such as the Riemannian manifold of SPD matrices, for some others, it is theoretically impossible.

In particular, if the manifold is compact, it is impossible to find an isometry between the manifold and \mathbb{R}^n , since \mathbb{R}^n is not compact. Therefore, it is not possible to obtain a positive definite Gaussian from the geodesic distance of a compact manifold. In such cases, the best hope is to find a different non-geodesic distance on the manifold that does not differ much from the geodesic distance, but still satisfies the conditions of Theorem 6.1.

6.3 Kernels on Sym_d^+

We now discuss different metrics on Sym_d^+ that can be used to define positive definite Gaussian kernels. Since Sym_d^+ is not compact, as explained in the previous section, there is some hope in finding a geodesic distance on it that also defines a positive definite Gaussian kernel. In this section we show that the log-Euclidean distance, which has been proved to be a geodesic distance on Sym_d^+ [31], is such a distance.

In the log-Euclidean framework, a geodesic connecting $S_1, S_2 \in \text{Sym}_d^+$ is defined as $\gamma(t) = \exp((1-t)\log(S_1) + t\log(S_2))$ for $t \in [0, 1]$. The log-Euclidean geodesic distance between S_1 and S_2 can be expressed as

$$d_{LE}(S_1, S_2) = \|\log(S_1) - \log(S_2)\|_F, \quad (1)$$

where $\|\cdot\|_F$ denotes the Frobenius matrix norm induced by the Frobenius matrix inner product $\langle \cdot, \cdot \rangle_F$.

The log-Euclidean distance has proven an effective distance measure on Sym_d^+ [10], [33]. Furthermore, it yields a positive definite Gaussian kernel as stated in the following corollary to Theorem 6.1:

Corollary 6.3 (Theorem 6.1). *The Log-Euclidean Gaussian kernel $k_{LE} : (\text{Sym}_d^+ \times \text{Sym}_d^+) \rightarrow \mathbb{R} : k_{LE}(S_1, S_2) := \exp(-\gamma d_{LE}^2(S_1, S_2))$, where $d_{LE}(S_1, S_2)$ is the log-Euclidean distance between S_1 and S_2 , is a positive definite kernel for all $\gamma > 0$.*

Proof. Directly follows from Theorem 6.1 with the Frobenius matrix inner product. \square

A number of other metrics have been proposed for Sym_d^+ [33]. The definitions and properties of these metrics are summarized in Table 1. Note that only some of them were derived by considering the Riemannian geometry of the manifold and hence define true geodesic distances. Similarly to the log-Euclidean metric, from Theorem 6.1, it

TABLE 1
Properties of Different Metrics on Sym_d^+

Metric Name	Formula	Geodesic Distance	Positive Definite Gaussian Kernel for all $\gamma > 0$
Log-Euclidean	$\ \log(\mathbf{S}_1) - \log(\mathbf{S}_2)\ _F$	Yes	Yes
Affine-Invariant	$\ \log(\mathbf{S}_1^{-1/2} \mathbf{S}_2 \mathbf{S}_1^{-1/2})\ _F$	Yes	No
Cholesky	$\ \text{chol}(\mathbf{S}_1) - \text{chol}(\mathbf{S}_2)\ _F$	No	Yes
Power-Euclidean	$\frac{1}{\alpha} \ \mathbf{S}_1^\alpha - \mathbf{S}_2^\alpha\ _F$	No	Yes
Root Stein Divergence	$[\log \det(\frac{1}{2} \mathbf{S}_1 + \frac{1}{2} \mathbf{S}_2) - \frac{1}{2} \log \det(\mathbf{S}_1 \mathbf{S}_2)]^{1/2}$	No	No

We analyze the positive definiteness of Gaussian kernels generated by different metrics. Theorem 6.1 applies to the metrics claimed to generate positive definite Gaussian kernels. For the other metrics, examples of non-positive definite Gaussian kernels exist.

directly follows that the Cholesky and power-Euclidean metrics also define positive definite Gaussian kernels for all values of γ . Note that some metrics may yield a positive definite Gaussian kernel for some value of γ only. This, for instance, was shown in [32] for the root Stein divergence metric. No such result is known for the affine-invariant metric. Constraints on γ are nonetheless undesirable, since one should be able to freely tune γ to reflect the data distribution, and automatic model selection algorithms require kernels to be positive definite for continuous values of $\gamma > 0$ [22].

6.4 Kernels on \mathcal{G}_n^r

Similarly to Sym_d^+ , different metrics can be defined on \mathcal{G}_n^r . Many of these metrics are related to the principal angles between two subspaces. Given two $n \times r$ matrices Y_1 and Y_2 with orthonormal columns, representing two points on \mathcal{G}_n^r , the principal angles between the corresponding subspaces are obtained from the singular value decomposition (SVD) of $Y_1^T Y_2$ [34]. More specifically, if USV^T is the singular value decomposition of $Y_1^T Y_2$, then the entries of the diagonal matrix S are the cosines of the principal angles between $[Y_1]$ and $[Y_2]$. Let $\{\theta_i\}_{i=1}^r$ represent those principal angles. Then, the geodesic distance derived from the canonical geometry of the Grassmann manifold, called the arc length, is given by $(\sum_i \theta_i^2)^{1/2}$ [34]. Unfortunately, as can be shown with a counter-example, this distance, when squared, is not negative definite and hence does not yield a positive definite Gaussian for all $\gamma > 0$. Given Theorem 6.2 and the discussion that followed, this not a surprising result: Since the Grassmann manifold is a compact manifold, it is not possible to find a geodesic distance that also yields a positive definite Gaussian for all $\gamma > 0$.

Nevertheless, there exists another widely used metric on the Grassmann manifold, namely the projection metric,

which gives rise to a positive definite Gaussian kernel. The projection distance between two subspaces $[Y_1], [Y_2]$ is given by

$$d_P([Y_1], [Y_2]) = 2^{-1/2} \|Y_1 Y_1^T - Y_2 Y_2^T\|_F. \quad (2)$$

We now formally introduce this Gaussian RBF kernel on the Grassmann manifold.

Corollary 6.4 (Theorem 6.1). *The Projection Gaussian kernel $k_P : (\mathcal{G}_n^r \times \mathcal{G}_n^r) \rightarrow \mathbb{R} : k_P([Y_1], [Y_2]) := \exp(-\gamma d_P^2([Y_1], [Y_2]))$, where $d_P([Y_1], [Y_2])$ is the projection distance between $[Y_1]$ and $[Y_2]$, is a positive definite kernel for all $\gamma > 0$.*

Proof. Follows from Theorem 6.1 with the Frobenius matrix inner product. \square

As shown in Table 2, none of the other popular metrics on Grassmann manifolds have this property. Counter-examples exist for these metrics to support our claims.

6.4.1 Calculation of the Projection Gaussian Kernel

To calculate the kernel introduced in Corollary 6.4, one needs to calculate the squared projection distance given by $d_P^2([Y_1], [Y_2]) = 2^{-1} \|Y_1 Y_1^T - Y_2 Y_2^T\|_F^2$ or $d_P^2([Y_1], [Y_2]) = (\sum_i \sin^2 \theta_i)^{1/2}$ where both Y_1 and Y_2 are $n \times r$ matrices. The second formula requires a singular value decomposition to find the θ_i s, which we would like to avoid due to its computational complexity. The first formula requires calculating the Frobenius l_2 distance between $Y_1 Y_1^T$ and $Y_2 Y_2^T$, both of which are $n \times n$ matrices. For many applications, n is quite large, and therefore, the direct computation of $\|Y_1 Y_1^T - Y_2 Y_2^T\|_F^2$ is inefficient. As a consequence, the cost of computing d_P^2 can be reduced with the following equation, which makes use of

TABLE 2
Properties of Different Metrics on \mathcal{G}_n^r

Metric Name	Formula	Geodesic Distance	Positive Definite Gaussian Kernel for all $\gamma > 0$
Projection	$2^{-1/2} \ Y_1 Y_1^T - Y_2 Y_2^T\ _F = (\sum_i \sin^2 \theta_i)^{1/2}$	No	Yes
Arc length	$(\sum_i \theta_i^2)^{1/2}$	Yes	No
Fubini-Study	$\arccos \det(Y_1^T Y_2) = \arccos(\prod_i \cos \theta_i)$	No	No
Chordal 2-norm	$\ Y_1 U - Y_2 V\ _2 = 2 \max_i \sin \frac{1}{2} \theta_i$	No	No
Chordal F-norm	$\ Y_1 U - Y_2 V\ _F = 2(\sum_i \sin^2 \frac{1}{2} \theta_i)^{1/2}$	No	No

Here, USV^T is the singular value decomposition of $Y_1^T Y_2$, whereas θ_i s are the principal angles between the two subspaces $[Y_1]$ and $[Y_2]$. Our claims on positive definiteness are supported by either Theorem 6.1, or counter-examples.

some properties of the matrix trace and of the fact that $Y_1^T Y_1 = Y_2^T Y_2 = I_r$:

$$d_P^2([Y_1], [Y_2]) = 2^{-1} \|Y_1 Y_1^T - Y_2 Y_2^T\|_F^2 = r - \|Y_1^T Y_2\|_F^2.$$

This implies that it is sufficient to compute only the Frobenius norm of $Y_1^T Y_2$, an $r \times r$ matrix, to calculate $d_P^2([Y_1], [Y_2])$.

7 KERNEL-BASED ALGORITHMS ON MANIFOLDS

A major advantage of being able to compute positive definite kernels on manifolds is that it directly allows us to make use of algorithms developed for \mathbb{R}^n , while still accounting for the geometry of the manifold. In this section, we discuss the use of five kernel-based algorithms on manifolds. The resulting algorithms can be thought of as generalizations of the original Euclidean kernel methods to manifolds. In the following, we use \mathcal{M} , $k(\cdot, \cdot)$, \mathcal{H} and $\phi(x)$ to denote a manifold, a positive definite kernel defined on $\mathcal{M} \times \mathcal{M}$, the RKHS generated by k , and the feature vector in \mathcal{H} to which $x \in \mathcal{M}$ is mapped, respectively. Although we use $\phi(x)$ for explanation purposes, following the kernel trick, it never needs to be explicitly computed in any of the algorithms.

7.1 Classification on Manifolds

We first consider the binary classification problem on a manifold. To this end, we propose to extend the popular Euclidean kernel SVM algorithm to manifold-valued data. Given a set of training examples $\{(x_i, y_i)\}_{i=1}^m$, where $x_i \in \mathcal{M}$ and the label $y_i \in \{-1, 1\}$, kernel SVM searches for a hyperplane in \mathcal{H} that separates the feature vectors of the positive and negative classes with maximum margin. The class of a test point $x \in \mathcal{M}$ is determined by the position of the feature vector $\phi(x)$ in \mathcal{H} relative to the separating hyperplane. Classification with kernel SVM can be done very fast, since it only requires to evaluate the kernel at the *support vectors*.

The above procedure is equivalent to solving the standard kernel SVM problem with kernel matrix generated by k . Thus, any existing SVM software package can be utilized for training and classification. Convergence of standard SVM optimization algorithms is guaranteed, since k is positive definite.

Kernel SVM on manifolds is much simpler to implement and less computationally demanding in both training and testing phases than the current state-of-the-art binary classification algorithms on manifolds, such as LogitBoost on a manifold [3], which involves iteratively combining weak learners on different tangent spaces. Weighted mean calculation in LogitBoost on a manifold involves an extremely expensive gradient descent procedure at each boosting iteration, which makes the algorithm scale poorly with the number of training samples. Furthermore, while LogitBoost learns classifiers on tangent spaces used as first order Euclidean approximations to the manifold, our approach works in a rich high dimensional feature space. As will be shown in our experiments, this yields better classification results.

With manifold-valued data, extending the current state-of-the-art binary classification methods to multi-class

classification is not straight-forward [19], [20]. By contrast, our manifold kernel SVM classification method can easily be extended to the multi-class case with standard one-vs-one or one-vs-all procedures.

7.2 Feature Selection on Manifolds

We next tackle the problem of combining multiple manifold-valued descriptors via a multiple kernel learning approach. The core idea of MKL is to combine kernels computed from different descriptors (e.g., image features) to obtain a kernel that optimally separates two classes for a given classifier. Here, we follow the formulation of [27] and make use of an SVM classifier. As a feature selection method, MKL has proven more effective than conventional techniques such as wrappers, filters and boosting [38].

More specifically, given training examples $\{(x_i, y_i)\}_1^m$, where $x_i \in \mathcal{X}$ (some nonempty set), $y_i \in \{-1, 1\}$, and a set of descriptor generating functions $\{g_j\}_1^N$ where $g_j: \mathcal{X} \rightarrow \mathcal{M}$, we seek to learn a binary classifier $f: \mathcal{X} \rightarrow \{-1, 1\}$ by selecting and optimally combining the different descriptors generated by g_1, \dots, g_N . Let $K^{(j)}$ be the kernel matrix generated by g_j and k as $K_{pq}^{(j)} = k(g_j(x_p), g_j(x_q))$. The combined kernel can be expressed as $K^* = \sum_j \lambda_j K^{(j)}$, where $\lambda_j \geq 0$ for all j guarantees the positive definiteness of K^* . The weights λ can be learned using a min-max optimization procedure with an l_1 regularizer on λ to obtain a sparse combination of kernels. The algorithm has two steps in each iteration: First it solves a conventional SVM problem with λ , hence K^* , fixed. Then it updates λ with the SVM parameters fixed. These two steps are repeated until convergence. For more details, we refer the reader to [27] and [38]. Note that convergence of MKL is only guaranteed if all the kernels are positive definite, which is satisfied in this setup since k is positive definite.

We also note that MKL on manifolds gives a convenient method to combine manifold-valued descriptors with Euclidean descriptors which is otherwise a difficult task due to their different geometries.

7.3 Dimensionality Reduction on Manifolds

We next study the extension of kernel PCA to nonlinear dimensionality reduction on manifolds. The usual Euclidean kernel PCA has proven successful in many applications [12], [39]. On a manifold, kernel PCA proceeds as follows: All points $x_i \in \mathcal{M}$ of a given dataset $\{x_i\}_{i=1}^m$ are mapped to feature vectors in \mathcal{H} , thus yielding the transformed set $\{\phi(x_i)\}_{i=1}^m$. The covariance matrix of this transformed set is then computed, which really amounts to computing the kernel matrix of the original data using the function k . An l -dimensional representation of the data is obtained by computing the eigenvectors of the kernel matrix. This representation can be thought of as a Euclidean representation of the original manifold-valued data. However, owing to our kernel, it was obtained by accounting for the geometry of \mathcal{M} .

Once the kernel matrix is calculated with k , implementation details of the algorithm are similar to that of the Euclidean kernel PCA algorithm. We refer the reader to [39] for more details on the implementation.

7.4 Clustering on Manifolds

For clustering problems on manifolds, we propose to make use of kernel k -means. Kernel k -means maps points to a high-dimensional Hilbert space and performs k -means clustering in the resulting feature space [39]. In a manifold setting, a given dataset $\{x_i\}_{i=1}^m$, with each $x_i \in \mathcal{M}$, is clustered into a pre-defined number of groups in \mathcal{H} , such that the sum of the squared distances from each $\phi(x_i)$ to the nearest cluster center is minimized. The resulting clusters can then act as classes for the $\{x_i\}_{i=1}^m$.

The unsupervised clustering method on Riemannian manifolds proposed in [5] clusters points in a low-dimensional space after dimensionality reduction on the manifold. In contrast, our method performs clustering in a high-dimensional RKHS which, intuitively, better represents the data distribution.

7.5 Discriminant Analysis on Manifolds

The kernelized version of linear discriminant analysis, known as kernel Fisher discriminant analysis (Kernel FDA), can also be extended to manifolds on which a positive definite kernel can be defined. Given $\{(x_i, y_i)\}_{i=1}^m$, with each $x_i \in \mathcal{M}$ having class label y_i , manifold kernel FDA maps each point x_i on the manifold to a feature vector in \mathcal{H} and finds a new basis in \mathcal{H} where the class separation is maximized. The output of the algorithm is a Euclidean representation of the original manifold-valued data, but with a larger separation between class means and a smaller within-class variance. Up to $(l - 1)$ dimensions can be extracted via kernel FDA where l is the number of classes. We refer the reader to [12], [40] for implementation details. In Euclidean spaces, kernel FDA has become an effective pre-processing step to perform nearest-neighbor classification in the highly discriminative, reduced dimensional space.

8 APPLICATIONS AND EXPERIMENTS

We now present two series of experiments on Sym_d^+ and \mathcal{G}_n^r using the positive definite kernels introduced in Section 6 and the algorithms described in Section 7.

8.1 Experiments on Sym_d^+

In the following, we use the log-Euclidean Gaussian kernel defined in Corollary 6.3 to apply different kernel methods to Sym_d^+ . We compare our kernel methods on Sym_d^+ to other state-of-the-art algorithms on Riemannian manifolds and to kernel methods on Sym_d^+ with the usual Euclidean Gaussian kernel that does not account for the nonlinearity of the manifold.

8.1.1 Pedestrian Detection

We first demonstrate the use of our log-Euclidean Gaussian kernel for the task of pedestrian detection with kernel SVM and MKL on Sym_d^+ . Let $\{(W_i, y_i)\}_{i=1}^m$ be the training set, where each $W_i \in \mathbb{R}^{h \times w}$ is an image window and $y_i \in \{-1, 1\}$ is the class label (background or person) of W_i . Following [3], we use covariance descriptors computed from the feature vector $[x, y, |I_x|, |I_y|, \sqrt{I_x^2 + I_y^2}, |I_{xx}|, |I_{yy}|, \arctan(\frac{|I_x|}{|I_y|})]$, where x, y are pixel locations and I_x, I_y, \dots are

intensity derivatives. The covariance matrix for an image patch of arbitrary size therefore is an 8×8 SPD matrix. In an $h \times w$ window W , a large number of covariance descriptors can be computed from subwindows with different sizes and positions sampled from W . We consider N subwindows $\{w_j\}_{j=1}^N$ of size ranging from $h/5 \times w/5$ to $h \times w$, positioned at all possible locations. The covariance descriptor of each subwindow is normalized using the covariance descriptor of the full window to improve robustness against illumination changes. Such covariance descriptors can be computed efficiently using integral images [3].

Let $X_i^{(j)} \in \text{Sym}_8^+$ denote the covariance descriptor of the j th subwindow of W_i . To reduce this large number of descriptors, we pick the best 100 subwindows that do not mutually overlap by more than 75 percent, by ranking them according to their variance across all training samples. The rationale behind this ranking is that a good descriptor should have a low variance across all given positive detection windows. Since the descriptors lie on a manifold, for each $X^{(j)}$, we compute a variance-like statistic $\text{var}(X^{(j)})$ across all positive training samples as

$$\text{var}(X^{(j)}) = \frac{1}{m_+} \sum_{i: y_i=1} d_g^p(X_i^{(j)}, \bar{X}^{(j)}), \quad (3)$$

where m_+ is the number of positive training samples and \bar{X} is the Karcher mean of $\{X_i\}_{i: y_i=1}$ given by $\bar{X} = \exp(\frac{1}{m_+} \sum_{i: y_i=1} \log(X_i))$ under the log-Euclidean metric. We set $p = 1$ in Eq. (3) to make the statistic less sensitive to outliers. We then use the SVM-MKL framework described in Section 7.2 to learn the final classifier, where each kernel is defined on one of the 100 selected subwindows. At test time, detection is achieved in a sliding window manner followed by a non-maximum suppression step.

To evaluate our approach, we made use of the INRIA person dataset [41]. Its training set consists of 2,416 positive windows and 1,280 person-free negative images, and its test set of 1,237 positive windows and 453 negative images. Negative windows are generated by sampling negative images [41]. We first used all positive samples and 12,800 negative samples (10 random windows from each negative image) to train an initial classifier. We used this classifier to find hard negative examples in the training images, and re-trained the classifier by adding these hard examples to the training set. Cross validation was used to determine the hyperparameters including the parameter γ of the kernels. We used the evaluation methodology of [41].

In Fig. 1, we compare the detection-error tradeoff (DET) curves of our approach and state-of-the-art methods. The curve for our method was generated by continuously varying the decision threshold of the final MKL classifier. We also evaluated our MKL framework with the Euclidean Gaussian kernel. Note that the proposed MKL method with our Riemannian kernel outperforms MKL with the Euclidean kernel, as well as LogitBoost on the manifold. This demonstrates the importance of accounting for the nonlinearity of the manifold using an appropriate positive definite kernel. It is also worth noting that LogitBoost on the manifold is significantly more complex and harder to implement than our method.

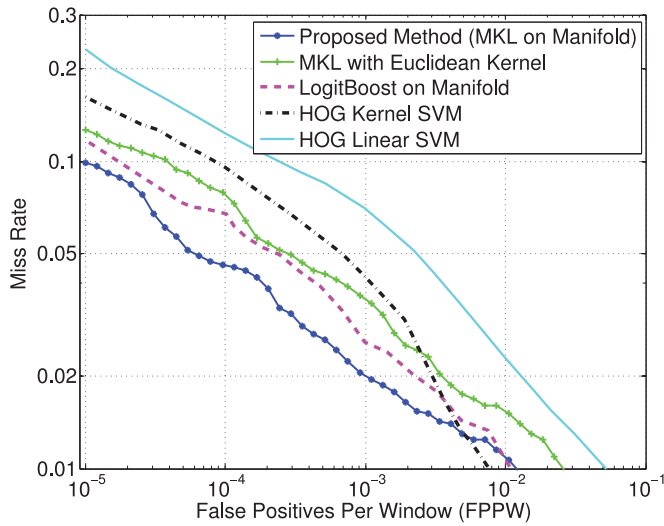


Fig. 1. Pedestrian detection. Detection-Error tradeoff curves for the proposed manifold MKL approach and state-of-the-art methods on the INRIA dataset. Our method outperforms existing manifold methods and Euclidean kernel methods. The curves for the baselines were reproduced from [3].

8.1.2 Visual Object Categorization

We next tackle the problem of unsupervised object categorization. To this end, we used the ETH-80 dataset [42] which contains eight categories with 10 objects each and 41 images per object. We used 21 randomly chosen images from each object to compute the parameter γ and the rest to evaluate clustering accuracy. For each image, we used a single 5×5 covariance descriptor calculated from the features $[x, y, I, |I_x|, |I_y|]$, where x, y are pixel locations and I, I_x, I_y are intensity and derivatives. To obtain object categories, the kernel k -means algorithm on Sym_5^+ described in Section 7.4 was employed.

One drawback of k -means and its kernel counterpart is their sensitivity to initialization. To overcome this, we ran each algorithm 20 times with random initializations and picked the iteration that converged to the minimum sum of point-to-centroid squared distances. For kernel k -means on Sym_5^+ , distances in the RKHS were used. We assumed the number of clusters to be known.

To set a benchmark, we evaluated the performance of both k -means and kernel k -means on Sym_5^+ with different metrics that generate positive definite Gaussian kernels (see Table 1). For the power-Euclidean metric, we used $\alpha = 0.5$, which, as in the non-kernel method of [33], we found to yield the best results. For all non-Euclidean metrics with

(non-kernel) k -means, the Karcher mean [33] was used to compute the centroid. The results of the different methods are summarized in Table 3. Manifold kernel k -means with the log-Euclidean Gaussian kernel performs significantly better than all other methods in all test cases. These results also outperform the results with the heat kernel reported in [24]. Note, however, that [24] only considered 3 and 4 classes without mentioning which classes were used.

8.1.3 Texture Recognition

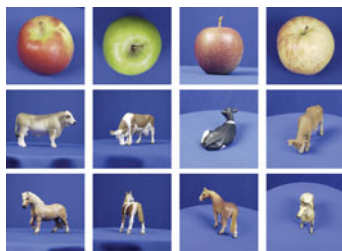
We then utilized our log-Euclidean Gaussian kernel to demonstrate the effectiveness of manifold kernel PCA on texture recognition. To this end, we used the Brodatz dataset [43], which consists of 111 different 640×640 texture images. Each image was divided into four subimages of equal size, two of which were used for training and the other two for testing.

For each training image, covariance descriptors of 50 randomly chosen 128×128 windows were computed from the feature vector $[I, |I_x|, |I_y|, |I_{xx}|, |I_{yy}|]$ [2]. Kernel PCA on Sym_5^+ with our Riemannian kernel was then used to extract the top l principal directions in the RKHS, and project the training data along those directions. Given a test image, we computed 100 covariance descriptors from random windows and projected them to the l principal directions obtained during training. Each such projection was classified using a majority vote over its five nearest-neighbors. The class of the test image was then decided by majority voting among the 100 descriptors. Cross validation on the training set was used to determine γ . For comparison purposes, we repeated the same procedure with the Euclidean Gaussian kernel. Results obtained for these kernels and different values of l are presented in Table 4. The better recognition accuracy indicates that kernel PCA with the Riemannian kernel more effectively captures the information of the manifold-valued descriptors than the Euclidean kernel.

8.1.4 Segmentation

We now illustrate the use of our kernel to segment different types of images. First, we consider DTI segmentation, which is a key application area of algorithms on Sym_d^+ . We utilized kernel k -means on Sym_3^+ with our Riemannian kernel to segment a real DTI image of the human brain. Each pixel of the input DTI image is a 3×3 SPD matrix, which can thus directly be used as input to the algorithm. The k clusters obtained by the algorithm act as classes, thus yielding a segmentation of the image.

TABLE 3
Object Categorization



Nb. of classes	Euclidean		Cholesky		Power-Euclidean		Log-Euclidean	
	KM	KKM	KM	KKM	KM	KKM	KM	KKM
3	72.50	79.00	73.17	82.67	71.33	84.33	75.00	94.83
4	64.88	73.75	69.50	84.62	69.50	83.50	73.00	87.50
5	54.80	70.30	70.80	82.40	70.20	82.40	74.60	85.90
6	50.42	69.00	59.83	73.58	59.42	73.17	66.50	74.50
7	42.57	68.86	50.36	69.79	50.14	69.71	59.64	73.14
8	40.19	68.00	53.81	69.44	54.62	68.44	58.31	71.44

Sample images and percentages of correct clustering on the ETH-80 dataset using k -means (KM) and kernel k -means (KKM) with different metrics on Sym_d^+ . The proposed KKM method with the log-Euclidean metric achieves the best results in all the tests.

TABLE 4
Texture Recognition

Kernel	Classification Accuracy			
	$l = 10$	$l = 11$	$l = 12$	$l = 15$
Log-Euclidean	95.50	95.95	96.40	96.40
Euclidean	89.64	90.09	90.99	91.89

Recognition accuracies on the Brodatz dataset with k -NN in an l -dimensional Euclidean space obtained by kernel PCA. The log-Euclidean Gaussian kernel introduced in this paper captures information more effectively than the usual Euclidean Gaussian kernel.

Fig. 2 depicts the resulting segmentation along with the ellipsoid and fractional anisotropy representations of the original DTI image. We also show the results obtained by replacing the Riemannian kernel with the Euclidean one. Note that, up to some noise due to the lack of spatial smoothing, Riemannian kernel k -means was able to correctly segment the corpus callosum from the rest of the image.

We then followed the same approach to perform 2D motion segmentation. To this end, we used a spatio-temporal structure tensor directly computed on image intensities (i.e., without extracting features such as optical flow). The spatio-temporal structure tensor for each pixel is computed as $T = K * (\nabla I \nabla I^T)$, where $\nabla I = (I_x, I_y, I_t)$ and $K*$ indicates convolution with the regular Gaussian kernel for smoothing purposes. Each pixel is thus represented as a 3×3 SPD matrix T and segmentation can be performed by clustering these matrices using kernel k -means on Sym_3^+ .

We applied this strategy to two images taken from the Hamburg Taxi sequence. Fig. 3 compares the results of kernel k -means with our log-Euclidean Gaussian kernel with the results of [5] obtained by first performing LLE, LE, or HLL on Sym_3^+ and then clustering in the low dimensional space. Note that our approach yields a much cleaner segmentation than the baselines. This could be attributed to the fact that we perform clustering in a high dimensional feature space, whereas the baselines work in a reduced dimensional space.

8.2 Experiments on \mathcal{G}_n^r

We now present our experimental evaluation of the proposed kernel methods on \mathcal{G}_n^r with the projection Gaussian kernel introduced in Corollary 6.4. We compare our results

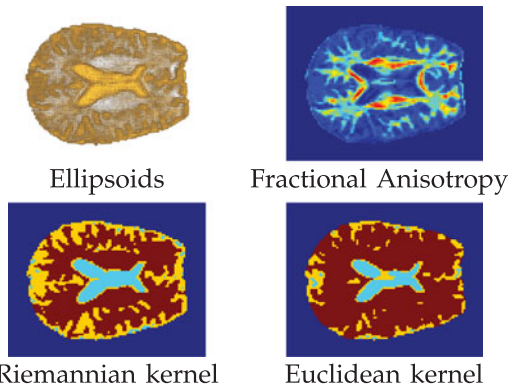


Fig. 2. DTI segmentation. Segmentation of the corpus callosum with kernel k -means on Sym_3^+ . The proposed kernel yields a cleaner segmentation.

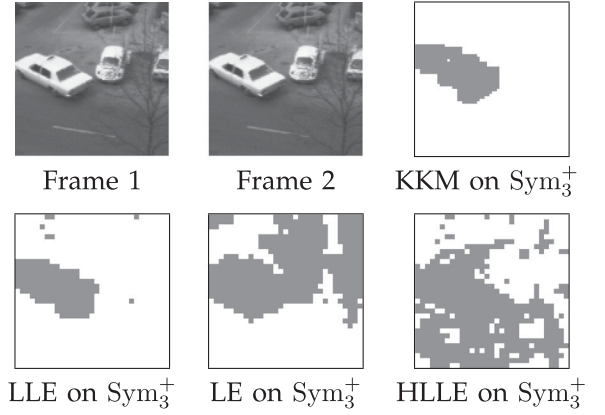


Fig. 3. 2D motion segmentation. Comparison of the segmentations obtained with kernel k -means with our Riemannian kernel (KKM), LLE, LE and HLL on Sym_3^+ . Our KKM algorithm yields a much cleaner segmentation. The baseline results were reproduced from [5].

to those obtained with state-of-the-art classification and clustering methods on \mathcal{G}_n^r , and show that we achieve significantly better classification and clustering accuracies in a number of different tasks.

In each of the following experiments, we model an image set with the linear subspace spanned by its principal components. More specifically, let $\{\mathbf{f}_i\}_{i=1}^p$, with each $\mathbf{f}_i \in \mathbb{R}^n$, be a set of descriptors each representing one image in a set of p images. The image set can then be represented by the linear subspace spanned by r ($r < p, n$) principal components of $\{\mathbf{f}_i\}_{i=1}^p$. Limiting r helps reducing noise and other fine variations within the image set, which are not useful for classification or clustering. The image set descriptors obtained in this manner are r -dimensional linear subspaces of the n -dimensional Euclidean space, which lie on the (n, r) Grassmann manifold \mathcal{G}_n^r . We note that the r -principal components of $\{\mathbf{f}_i\}_{i=1}^p$ can be efficiently obtained by performing a Singular Value Decomposition on the $n \times p$ matrix F having the \mathbf{f}_i s as columns: If USV^T is the singular value decomposition of F , the columns of U corresponding to the largest r singular values give the r principal components of $\{\mathbf{f}_i\}_{i=1}^p$.

8.2.1 Video Based Face Recognition

Face recognition from video, which uses an image set for identification of a person, is a rapidly developing area in computer vision. For this task, the videos are often modeled as linear subspaces, which lie on a Grassmann manifold [7], [8]. To demonstrate the use of our projection Gaussian kernel in video based face recognition, we used the YouTube Celebrity dataset [44], which contains 1,910 video clips of 47 different people. Face recognition on this dataset is challenging since the videos are highly compressed and most of them have low resolution.

We used the cascaded face detector of [45] to extract face regions from videos and resized them to have a common size of 96×96 . Each face image was then represented by a histogram of Local Binary Patterns [46] having 232 equally-spaced bins. We next represented each image set corresponding to a single video clip by a linear subspace of order 5. We randomly chose 70 percent of the dataset for training and the remaining 30 percent for testing. We report the classification accuracy averaged over 10 different random splits.

TABLE 5
Face and Action Recognition

Method	Face Recognition Accuracy	Action Recognition Accuracy
DCC [47]	60.21 \pm 2.9	41.95 \pm 9.6
KAHM [48]	67.49 \pm 3.5	70.05 \pm 0.9
GDA [7]	58.72 \pm 3.0	67.33 \pm 1.1
GGDA [8]	61.05 \pm 2.2	73.54 \pm 2.0
Linear SVM	64.76 \pm 2.1	74.66 \pm 1.2
Manifold Kernel FDA	65.32 \pm 1.4	76.35 \pm 1.0
Manifold Kernel SVM	71.78 \pm 2.4	76.95 \pm 0.9

Recognition accuracies on the YouTube Celebrity and Ballet datasets. Our manifold kernel SVM method achieves the best results.

We employed both kernel SVM on a manifold and kernel FDA on a manifold with our projection Gaussian kernel. With kernel SVM, a one-vs-all procedure was used for multiclass classification. For kernel FDA, the training data was projected to an $(l - 1)$ dimensional space, where $l = 47$ is the number of classes, and we used a 1-nearest-neighbor method to predict the class of a test sample projected to the same space. We determined the hyperparameters of both methods using cross-validation on the training data.

We compared our approach with several state-of-the-art image set classification methods: Discriminant analysis of Canonical Correlations (DCC) [47], Kernel Affine Hull Method (KAHM) [48], Grassmann Discriminant Analysis (GDA) [7], and Graph-embedding Grassmann Discriminant Analysis (GGDA) [8]. As shown in Table 5, manifold kernel SVM achieves the best accuracy. GDA uses the Grassmann projection kernel which corresponds to the linear kernel with FDA in a Euclidean space. Therefore, the results of GDA and Manifold Kernel FDA in Table 5 also provide a nice comparison between the linear kernel and our projection Gaussian kernel. To obtain a similar comparison with SVMs, we also performed one-vs-all SVM classification with the projection kernel, which really amounts to linear SVM in the Projective space. This method is denoted by Linear SVM in Table 5. With both FDA and SVM, our Gaussian kernel performs better than the projection kernel, thus agreeing with the observation in Euclidean spaces that the Gaussian kernel performs better than the linear kernel.

8.2.2 Action Recognition

We next demonstrate the benefits of our projection Gaussian kernel on action recognition. To this end, we used the Ballet dataset [49], which contains 44 videos of eight actions performed by three different actors. Each video contains different actions performed by the same actor. Action recognition on this dataset is challenging due to large intra-class variations in clothing and movement.

We grouped every six subsequent frames containing the same action, which resulted in 2,338 image sets. Each frame was described by a Histogram of Oriented Gradients (HOG) descriptor [41], and four principal components were used to represent each image set. The samples were randomly split into two equally-sized sets to obtain training and test data. We report the average accuracy over 10 different splits.

As in the previous experiment, we used kernel FDA and one-vs-all SVM with the projection Gaussian kernel and

compared our methods with DCC, KAHM, GDA, GGDA, and Linear SVM. As evidenced by the results in Table 5, Manifold Kernel FDA and Manifold Kernel SVM both achieve superior performance compared to the state-of-the-art algorithms. Out of the two methods proposed in this paper, Manifold Kernel SVM achieves the highest accuracy.

8.2.3 Pose Categorization

Finally, we demonstrate the use of our projection Gaussian kernel in clustering on the Grassmann manifold. To this end, we used the well-known CMU-PIE face dataset [50], which contains face images of 67 subjects with 13 different poses and 21 different illuminations. Images were closely cropped to enclose the face region followed by a resizing to 32×32 . The vectorized intensity values were directly used to describe each image. We used images of the same subject with the same pose but different illuminations to form an image set, which was then represented by a linear subspace of order 6. This resulted in a total of $67 \times 13 = 871$ Grassmann descriptors, each lying on \mathcal{G}_{1024}^6 .

The goal of the experiment was to cluster together image sets having the same pose. We randomly divided the images of the same subject with the same pose into two equally sized sets to obtain two collections of 871 image sets. The optimum value for γ was determined with the first collection and the results are reported on the second one. We compare our Manifold Kernel k -means (MKKM) with two other algorithms on the same data. The first algorithm, proposed in [9], is the conventional k -means with the arc length distance on \mathcal{G}_n^r and the corresponding Karcher mean (KM-AL). The publicly available code of [9] was used to obtain the results with KM-AL. Since the Karcher mean with the arc length distance does not have a closed-form solution and has to be calculated using a gradient descent procedure, KM-AL tends to slow down with the dimensionality of the Grassmann descriptors and the number of samples. The second baseline was k -means with the projection metric and the corresponding Karcher mean (KM-PM). Although the Karcher mean can be calculated in closed-form in this case, the algorithm becomes slow when working with large matrices. In our setup, since the projection space consisted of symmetric matrices of size $1,024 \times 1,024$, projection k -means boils down to performing k -means in a $1,024 \times (1,024 + 1)/2 = 524,800$ dimensional space.

The results of the three clustering algorithms for different numbers of clusters are given in Table 6. As in the experiment on Sym_d^+ , each clustering algorithm was run 20 times with different random initializations, and the iteration which converged to the minimum energy was picked. Note that our MKKM algorithm yields the best performance in all scenarios. Performance gain with the MKKM algorithm becomes more significant when the problem gets harder with more clusters.

9 CONCLUSION

In this paper, we have introduced a unified framework to analyze the positive definiteness of the Gaussian RBF kernel defined on a manifold or a more general metric space. We have then used the same framework to derive provably

TABLE 6
Pose Grouping

Nb. of Clusters	Clustering Accuracy		
	KM-AL [9]	KM-PM	MKKM
5	88.06	94.62	96.12
6	85.07	94.52	95.27
7	85.50	94.88	95.52
8	85.63	90.93	95.34
9	73.96	79.10	83.08
10	70.30	78.95	81.79
11	68.38	78.56	81.41
12	64.55	74.75	81.22
13	61.65	73.82	80.14

Clustering accuracies on the CMU-PIE dataset. The proposed kernel k -means method yields the best results in all the test cases.

positive definite kernels on the Riemannian manifold of SPD matrices and on the Grassmann manifold. These kernels were then utilized to extend popular learning algorithms designed for Euclidean spaces, such as SVM and FDA, to manifolds. Our experimental evaluation on several challenging computer vision tasks, such as pedestrian detection, object categorization, segmentation, action recognition and face recognition, has evidenced that identifying positive definite kernel functions on manifolds can be greatly beneficial when working with manifold-valued data. In the future, we intend to study this problem for other nonlinear manifolds, as well as to extend our theory to more general kernels than the Gaussian RBF kernel.

ACKNOWLEDGMENTS

This work was supported in part by an ARC grant. The authors would like to thank Bob Williamson for useful discussions. NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council (ARC) through the ICT Centre of Excellence program. S. Jayasumana is the corresponding author.

REFERENCES

- [1] X. Pennec, P. Fillard, and N. Ayache, "A Riemannian framework for tensor computing," *Int. J. Comput. Vis.*, vol. 66, no. 1, pp. 41–66, 2006.
- [2] O. Tuzel, F. Porikli, and P. Meer, "Region covariance: A fast descriptor for detection and classification," in *Proc. 9th Eur. Conf. Comput. Vis.*, 2006, pp. 589–600.
- [3] O. Tuzel, F. Porikli, and P. Meer, "Pedestrian detection via classification on Riemannian manifolds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 10, pp. 1713–1727, Oct. 2008.
- [4] P. Li and Q. Wang, "Local log-Euclidean covariance matrix (L2ECM) for image representation and its applications," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 469–482.
- [5] A. Goh and R. Vidal, "Clustering and dimensionality reduction on Riemannian manifolds," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–7.
- [6] R. Caseiro, J. F. Henriques, P. Martins, and J. Batista, "A nonparametric Riemannian framework on tensor field with application to foreground segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 1–8.
- [7] J. Hamm and D. D. Lee, "Grassmann discriminant analysis: A unifying view on subspace-based learning," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008.
- [8] M. T. Harandi, C. Sanderson, S. Shirazi, and B. C. Lovell, "Graph embedding discriminant analysis on Grassmannian manifolds for improved image set matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 2705–2712.
- [9] P. Turaga, A. Veeraraghavan, A. Srivastava, and R. Chellappa, "Statistical computations on Grassmann and Stiefel manifolds for image and video-based recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2273–2286, Nov. 2011.
- [10] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache, "Log-Euclidean metrics for fast and simple calculus on diffusion tensors," *Magn. Resonance Med.*, vol. 56, pp. 411–421, 2006.
- [11] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
- [12] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2002.
- [13] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu, "Semantic segmentation with second-order pooling," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 430–443.
- [14] P. Li, Q. Wang, W. Zuo, and L. Zhang, "Log-Euclidean kernels for sparse representation and dictionary learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1601–1608.
- [15] M. Harandi, C. Sanderson, R. Hartley, and B. Lovell, "Sparse coding and dictionary learning for symmetric positive definite matrices: A kernel approach," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 216–229.
- [16] D. Tosato, M. Spera, M. Cristani, and V. Murino, "Characterizing humans on Riemannian manifolds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1972–1984, Aug. 2013.
- [17] J. Malcolm, Y. Rath, and A. Tannenbaum, "A graph cut approach to image segmentation in tensor space," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.
- [18] Y. M. Lui, "Advances in matrix manifolds for computer vision," *Image Vision Comput.*, vol. 30, pp. 380–388, 2012.
- [19] D. Tosato, M. Farenzena, M. Cristani, M. Spera, and V. Murino, "Multi-class classification on Riemannian manifolds for video surveillance," in *Proc. 11th Eur. Conf. Comput. Vis.*, 2010, pp. 378–391.
- [20] R. Caseiro, P. Martins, J. F. Henriques, F. S. Leite, and J. Batista, "Rolling Riemannian manifolds to solve the multi-class classification problem," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 41–48.
- [21] M. Harandi, C. Sanderson, A. Wiliem, and B. Lovell, "Kernel analysis over Riemannian manifolds for visual recognition of actions, pedestrians and textures," in *Proc. IEEE Workshop Appl. Comput. Vis.*, 2012, pp. 433–439.
- [22] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, "Choosing multiple parameters for support vector machines," *Mach. Learn.*, vol. 46, pp. 131–159, 2002.
- [23] J. Hamm and D. Lee, "Extended Grassmann kernels for subspace-based learning," in *Proc. Adv. Neural Inform. Process. Syst.*, 2008, pp. 601–608.
- [24] R. Caseiro, J. F. Henriques, P. Martins, and J. Batista, "Semi-intrinsic mean shift on Riemannian manifolds," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 342–355.
- [25] R. Vemulapalli, J. K. Pillai, and R. Chellappa, "Kernel learning for extrinsic classification of manifold features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 1782–1789.
- [26] J. Platt, "Fast training of support vector machines using sequential minimal optimization," in *Advances in Kernel Methods—Support Vector Learning*. Cambridge, MA, USA: MIT Press, 1999.
- [27] M. Varma and D. Ray, "Learning the discriminative power-invariance trade-off," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, 2007, pp. 1–8.
- [28] C. Ong, X. Mary, S. Canu, and S. A. J. Smola, "Learning with non-positive kernels," in *Proc. Int. Conf. Mach. Learn.*, 2004, pp. 639–646.
- [29] G. Wu, E. Y. Chang, and Z. Zhang, "An analysis of transformation on non-positive semi-definite similarity matrix for kernel machines," UCSB Technical Report, University of California, Santa Barbara (UCSB), California, USA, Apr. 2005.
- [30] M. Belkin and P. Niyogi, "Semi-supervised learning on Riemannian manifolds," *Mach. Learn.*, vol. 56, pp. 209–239, 2004.
- [31] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache, "Fast and simple computations on tensors with log-Euclidean metrics," INRIA Sophia Antipolis-Méditerranée 2004 route des lucioles-BP 93 06902 Sophia Antipolis Cedex France, Res. Rep. RR-5584, 2005.

- [32] S. Sra, "Positive definite matrices and the symmetric stein divergence," *Preprint*, 2012, [arXiv:1110.1773].
- [33] I. Dryden, A. Koloydenko, and D. Zhou, "Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging," *Ann. Appl. Statist.*, vol. 3, no. 3, pp. 1102–1123, 2009.
- [34] A. Edelman, T. A. Arias, and S. T. Smith, "The geometry of algorithms with orthogonality constraints," *SIAM J. Matrix Anal. Appl.*, vol. 20, pp. 303–353, 1999.
- [35] N. Aronszajn, "Theory of reproducing kernels," *Trans. Amer. Math. Soc.*, vol. 68, pp. 337–404, 1950.
- [36] C. Berg, J. P. R. Christensen, and P. Ressel, *Harmonic Analysis on Semigroups*. New York, NY, USA: Springer, 1984.
- [37] I. J. Schoenberg, "Metric spaces and positive definite functions," *Trans. Amer. Math. Soc.*, vol. 44, pp. 522–536, 1938.
- [38] M. Varma and B. R. Babu, "More generality in efficient multiple kernel learning," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 1065–1072.
- [39] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, pp. 1299–1319, 1998.
- [40] S. Mika, G. Ratsch, J. Weston, B. Schölkopf, and K. R. Mullers, "Fisher discriminant analysis with kernels," in *Proc. Workshop Neural Netw. Signal Process.*, 1999, pp. 41–48.
- [41] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 886–893.
- [42] B. Leibe and B. Schiele, "Analyzing appearance and contour based methods for object categorization," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit.*, 2003, pp. 409–415.
- [43] T. Randen and J. Husoy, "Filtering for texture classification: A comparative study," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 4, pp. 291–310, Apr. 1999.
- [44] M. Kim, S. Kumar, V. Pavlovic, and H. A. Rowley, "Face tracking and recognition with visual constraints in real-world videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.
- [45] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, pp. 137–154, 2004.
- [46] T. Ojala, M. Pietikainen, and T. Maenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.
- [47] T.-K. Kim, J. Kittler, and R. Cipolla, "Discriminative learning and recognition of image set classes using canonical correlations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 1005–1018, Jun. 2007.
- [48] H. Cevikalp and B. Triggs, "Face recognition based on image sets," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 2567–2573.
- [49] Y. Wang and G. Mori, "Human action recognition by semilabelled topic models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 10, pp. 1762–1774, Oct. 2009.
- [50] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression (PIE) database," in *Proc. 5th IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2002, pp. 46–51.



Sadeep Jayasumana received the BSc degree in electronics and telecommunication engineering from the University of Moratuwa, Sri Lanka, in 2010. He is now working towards the PhD degree at the College of Engineering and Computer Science, Australian National University (ANU). He is also a member of the Computer Vision Research Group at NICTA, government funded research laboratory. He received CSiRA Best Recognition prize at DICTA 2013. He is a student member of the IEEE.

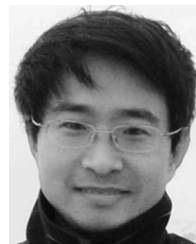


the book *Multiple View Geometry in Computer Vision*. He is a fellow of the IEEE.

Richard Hartley is a member of the computer vision group in the Research School of Engineering, at ANU, where he has been since January, 2001. He is also a member of the computer vision research group in NICTA. He worked at the GE Research and Development Center from 1985 to 2001, working first in VLSI design, and later in computer vision. He became involved with Image Understanding and Scene Reconstruction working with GE's Simulation and Control Systems Division. He is an author (with A. Zisserman) of the book *Multiple View Geometry in Computer Vision*. He is a fellow of the IEEE.



Mathieu Salzmann received the MSc and PhD degrees from EPFL in 2004 and 2009, respectively. He then joined the International Computer Science Institute and the EECS Department at the University of California at Berkeley as a post-doctoral fellow, and later the Toyota Technical Institute at Chicago as a research assistant professor. He is now a senior researcher at NICTA in Canberra. He is a member of the IEEE.



ber of ARC Centre of Excellence for Robotic Vision, a former member of the "Bionic Vision Australia", and a member of the IEEE.

Hongdong Li is with the computer vision group at ANU. His research interests include 3D Computer Vision, vision geometry, image and pattern recognition, and mathematical optimization. Prior to 2010, he was a senior research scientist with the NICTA, and a fellow at the RSISE, ANU. He received the CVPR Best Paper Award in 2012, Best Student Paper Award at ICPR 2010, CSiRA Best Recognition Paper Prize at DICTA 2013. He was in program committees (area chair/reviewer) for recent ICCV, CVPR, and ECCV. He is a member of ARC Centre of Excellence for Robotic Vision, a former member of the "Bionic Vision Australia", and a member of the IEEE.



Mehrtash Harandi received the BSc degree in electronics from the Sharif University of technology, MSc and PhD degrees in computer science from the University of Tehran, Iran. He is a senior researcher at the Computer Vision Research Group (CVRG), NICTA. His main research interests are theoretical and computational methods in computer vision and machine learning with a focus on Riemannian geometry. He is a member of the IEEE.

► For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/publications/dlib.