Combining Multiple Manifold-valued Descriptors for Improved Object Recognition

Sadeep Jayasumana\(^1\), Richard Hartley\(^1,2\), Mathieu Salzmann\(^2\), Hongdong Li\(^1\), and Mehrtash Harandi\(^2\)
\(^1\)Australian National University, Canberra \(^2\)NICTA, Canberra
sadeep.jayasumana@anu.edu.au

Abstract—We present a learning method for classification using multiple manifold-valued features. Manifold techniques are becoming increasingly popular in computer vision since Riemannian geometry often comes up as a natural model for many descriptors encountered in different branches of computer vision. We propose a feature combination and selection method that optimally combines descriptors lying on different manifolds while respecting the Riemannian geometry of each underlying manifold. We use our method to improve object recognition by combining HOG [1] and Region Covariance [2] descriptors that reside on two different manifolds. To this end, we propose a kernel on the \(n\)-dimensional unit sphere and prove its positive definiteness. Our experimental evaluation shows that combining these two powerful descriptors using our method results in significant improvements in recognition accuracy.

I. INTRODUCTION

Nonlinear data that lack a vector space structure are commonly encountered in many branches of computer vision. Examples include normalized histogram vectors [1], [3] and covariance descriptors [2, 4] found in object detection/recognition, diffusion tensors in biomedical image analysis [5], [6], 3D rotation matrices in geometrical computer vision [7] and linear subspaces of the \(n\)-dimensional Euclidean space in video based vision [8]. The spaces where such nonlinear data lie lack a vector space structure in the sense that they are either not closed under vector addition and scalar multiplication, or these operations are not defined in them at all. For example, \(d \times d\) covariance matrices, commonly used as region descriptors in object detection, form a convex cone in the \(d(d + 1)/2\) Euclidean space. This convex cone is not closed under scalar multiplication.

Although the nonlinear data classes stated above do not have a vector space structure and do not adhere to Euclidean geometry, many of them do possess interesting geometries that are studied under a separate branch of mathematics: Riemannian geometry. Riemannian geometry provides tools to extend some Euclidean notions such as inner products and angles between curves, to nonlinear manifolds. Conventionally, computer vision and machine learning algorithms are developed for Euclidean spaces assuming linear (vector space) structure of the data. Utilizing these Euclidean techniques on manifold-valued data is not always straightforward. However, when the nonlinear data at hand lie on a Riemannian manifold, some Euclidean methods can be generalized to the manifold-valued data using tools provided by Riemannian geometry.

A common technique used to generalize Euclidean algorithms to Riemannian manifolds is to first obtain a Euclidean representation of the manifold-valued data by approximating the manifold by the tangent space at some point (usually the sample mean) on the manifold [4]. However, this technique only gives a first order approximation of the manifold and hence results in significant distortion of the original data distribution, specially in the areas far away from the point whose tangent space is used. Moreover, the extensive use of Riemannian tools such as exponential maps and logarithmic maps in such algorithms makes them inefficient.

An alternative approach to generalizing Euclidean algorithms to a given manifold is to embed the manifold in a high dimensional Reproducing Kernel Hilbert Space (RKHS) using a positive definite kernel defined on the manifold. This method has drawn significant attention in recent years [9]–[12]. In this approach, the manifold is embedded in a linear Hilbert space, making it possible to utilize Euclidean methods on manifold-valued data, while simultaneously obtaining a richer, higher dimensional representation of the original data distribution. This approach has shown to perform better than tangent space methods in many instances [10], [12].

In this paper, we address the problem of combining multiple manifold-valued descriptors for improved object recognition. It is well known that using more than one descriptor in a feature selection framework enhances the recognition/detection accuracy significantly [4], [13], [14]. For vector-valued descriptors, a number of feature combination and selection methods, ranging from simple concatenation of multiple features into a single vector to boosting, are available and commonly used. However, combining multiple features lying on different manifolds while respecting the true geometries of the underlying manifolds is not straightforward and has received little attention.

Here, we propose a method for combining multiple manifold-valued descriptors via RKHS embedding. We use positive definite kernels defined on the manifolds that account for their true geometries to embed the manifolds in Hilbert spaces and combine features in those Hilbert spaces. As a concrete example, we consider two specific manifolds, the unit \(n\)-sphere \(S^n\) in the \(n+1\) dimensional Euclidean space and the Riemannian manifold of \(d \times d\) Symmetric Positive Definite (SPD) matrices \(\text{Sym}_{++}^d\), and show how our method can be used to combine descriptors sampled from these two manifolds.

Contributions: The present paper makes two main contributions: First, it proposes a new, provably positive definite kernel on \(S^n\) which, unlike the usual Gaussian RBF kernel, accounts for the true geometry of the sphere. Second, this paper introduces a method to optimally combine two well-known
and extremely successful region descriptors, namely Histogram of Oriented Gradients (HOG) \[1\] and Region Covariance descriptors \[2, 4\], in order to improve object recognition. These two descriptor types lie on two different Riemannian manifolds. Our method optimally combines the two manifold-valued descriptors in order to maximize the object recognition accuracy. It will be shown in our experiments that optimally combining these two powerful region descriptors results in significant improvements in object recognition.

II. RELATED WORK

A significant amount of research has been done in recent years on generalizing Euclidean computer vision and machine learning techniques to Riemannian manifolds. This includes works in binary classification on a manifold \[4\], multi-class classification on a manifold \[9, 15\], clustering \[8\], dimensionality reduction \[12\] and interpolation \[6\]. Most of these works focus on a single manifold with a specific geometry. For example, \[4, 6\] consider the Riemannian manifold of SPD matrices, \[8, 9\] consider the Grassmann manifold. Combining or selecting features lying on different manifolds has received very little attention.

Two specific Riemannian manifolds encountered very often in computer vision are the unit \(n\)-sphere \(S^n\) and the Riemannian manifold of SPD matrices \(\text{Sym}_+^n\). Some examples for descriptors lying on \(S^n\) are famous SIFT descriptors \[3\], HOG descriptors \[1\]. Local Binary Patterns (LBP) descriptors \[16\] and any histogram representation in general which is subjected to direct or block \(l^2\) normalization. Examples of descriptors sampled from \(\text{Sym}_+^n\) include Region Covariance descriptors \[2\], diffusion tensors \[6\] and structure tensors \[12\]. Kernels on \(\text{Sym}_+^n\) that account for the Riemannian geometry of the manifold have been proposed in \[12\]. However, for descriptors lying on \(S^n\), which also is a Riemannian manifold, the traditional Euclidean Gaussian RBF is usually employed \[1\], neglecting the true geometry of the manifold.

Histogram of Oriented Gradients (HOG) descriptors were first proposed in \[1\] for human detection and have subsequently become very popular as region descriptors for object classification and detection. After the mandatory block normalization step, HOG descriptors lie on the \(n\)-dimensional sphere of some fixed radius, whose geometry is the same as that of the unit sphere \(S^n\). As an alternative region/object descriptor, Region Covariance descriptors first emerged in \[2\] and thereafter found applications in texture recognition \[17\], face recognition \[17\], action recognition \[18\] and tracking \[18\]. Covariance descriptors, being SPD matrices, lie on the Riemannian manifold \(\text{Sym}_+^n\). It has been shown in many occasions that accounting for the geometry of \(\text{Sym}_+^n\) is key to the success of algorithms operating on covariance descriptors \[4, 10, 12\].

Kernel methods are extensively used in Euclidean spaces mainly for classification with SVM and also for clustering, dimensionality reduction and regression \[19, 20\]. In recent years, there have been a series of works targeting the generalization of kernel methods to Riemannian manifolds \[9–12\]. The main challenge in generalizing kernel methods to a given Riemannian manifold lies in defining a kernel on the manifold that encodes the nonlinear geometry of the manifold while being positive definite. According to the Mercer’s theorem \[19\] only a positive definite kernel yields a valid embedding in an RKHS. Moreover, the positive definiteness of the kernel(s) being used is a requirement for the convergence of many popular learning algorithms \[14, 21\].

III. MANIFOLDS AND KERNELS

In this section, we briefly review the two manifolds used in the paper and positive definite kernels defined on them that permits us to embed the manifold under consideration in a high dimensional Hilbert space.

In differential geometry, a topological manifold, also known simply as a manifold, is a topological space (a set with the notion of neighborhood or open sets) which is locally similar to some Euclidean space. A differentiable manifold is a topological manifold equipped with a globally defined differential structure that allows one to perform calculus on the manifold. Finally, a Riemannian manifold is defined as a differentiable manifold with a smoothly varying inner product defined on the tangent bundle.

The geodesic between two points on a Riemannian manifold can be thought of as the shortest curve connecting the two points without leaving the manifold. Geodesics correspond to straight lines in Euclidean spaces. Therefore, the length of the connecting geodesic, dubbed geodesic distance, is the most suitable distance measure between two points lying on a Riemannian manifold.

A. The Unit \(n\)-sphere

The \(n\)-dimensional sphere that has unit radius and is centered at the origin of the \(n+1\) dimensional Euclidean space, denoted by \(S^n\), is perhaps the simplest Riemannian manifold after the Euclidean space itself. It inherits a Riemannian metric from its embedding in \(\mathbb{R}^{n+1}\). Under this Riemannian metric, the geodesic distance \(d_g\) between two points \(x, y \in S^n\) is simply the great circle distance between the two points, which is defined formally as

\[
d_g(x, y) = \arccos(x^T y),
\]

where \(\arccos : [-1, 1] \to [0, \pi]\) is the usual inverse cosine function.

Almost every descriptor used in computer vision that is derived from a histogram is ultimately normalized, either fully or block-wise, using the \(l^2\) norm \[1, 3, 16\]. The resulting descriptors therefore lie on \(S^n\), for some \(n\). When block normalization is used, the radius of the sphere might not be unit, but since any \(n\)-dimensional sphere centered at the origin is homeomorphic to \(S^n\), their geometries turn out to be exactly the same. For simplicity, one can think of this as scaling the block-normalized vectors by a constant, which does not alter the data in any way.

Although the actual geometry of \(S^n\) is not Euclidean, conventionally only Euclidean kernels, such as the linear kernel and the Gaussian RBF kernel with the Euclidean distance, have been used to perform kernel methods on descriptors lying on \(S^n\) \[1\]. In this paper, we propose a provably positive definite kernel that is derived from the geodesic distance on \(S^n\). This proposed kernel, named geodesic exponential kernel, permits
us to embed $S^n$ in a linear RKHS while accounting for the true geometry of the manifold.

B. The Riemannian manifold of SPD matrices

Symmetric Positive Definite (SPD) matrices are characterized by the property that their eigenvalues are positive. It has been shown that the space of $d \times d$ SPD matrices, which we denote by $Sym^+_d$, form a Riemannian manifold when endowed with an appropriate metric $[6]$, $[22]$. Popular Riemannian metrics for this space are the Affine-invariant metric $[6]$ and the Log-Euclidean metric $[3]$, $[22]$. Under the Log-Euclidean framework, the geodesic distance $d_{LE}$ between two SPD matrices $S_1$, $S_2 \in Sym^+_d$ is given by

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

where $\log(.)$ denotes the ordinary matrix logarithm which is well-defined for SPD matrices. Recently in $[22]$, it was proven in that the Gaussian RBF kernel defined with this distance is positive definite and performs better than the usual Euclidean Gaussian kernel. We use this kernel, named the log-Euclidean Gaussian kernel, in this paper, to embed the Gaussian kernel. We use this kernel, named the log-Euclidean Gaussian kernel, in this paper, to embed the Gaussian kernel. We use this kernel, named the log-Euclidean Gaussian kernel, in this paper, to embed the Gaussian kernel.

Theorem 1. The geodesic exponential kernel $k_g(x, y) : S^n \times S^n \to \mathbb{R}$ defined by $k_g(x, y) = \exp(-d_g(x, y)/2\sigma^2)$, where $d_g$ is the geodesic distance on $S^n$, is a positive definite kernel for all $\sigma$.

Proof: We present the proof of this theorem in a number of steps, starting from the following definition of positive and negative definite kernels $[23]$.

Definition 2. Let $X$ be a nonempty set. A function $f : (X \times X) \to \mathbb{R}$ is called a positive (resp. negative) definite kernel if and only if $f$ is symmetric (i.e., $f(x, y) = f(y, x)$ for all $x, y \in X$) and

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

for all $m \in \mathbb{N}$, $\{x_1, \ldots, x_m\} \subseteq X$ and $\{c_1, \ldots, c_m\} \subseteq \mathbb{R}$ (resp. $\sum_{i=1}^{m} c_i = 0$ in addition).

We then prove the following lemma using the above definition.

Lemma 3. Let $f : (X \times X) \to \mathbb{R}$ be a positive definite kernel defined on some nonempty set $X$. Then, $g(x, y) = \lambda - f(x, y)$ is negative definite for all $\lambda \in \mathbb{R}$.

Proof: Since $f$ is symmetric, so is $g$. We then need to prove that $\sum_{i,j=1}^{m} c_i c_j g(x_i, x_j) \leq 0$ for all $m \in \mathbb{N}$, $\{x_1, \ldots, x_m\} \subseteq X$ and $\{c_1, \ldots, c_m\} \subseteq \mathbb{R}$ with $\sum_{i=1}^{m} c_i = 0$.

We also use the following two theorems proved in Chapter 3 of $[23]$ to obtain the proof of Theorem

Theorem 4. Let $f : (X \times X) \to \mathbb{R}$ be a kernel defined on some nonempty set $X$. The kernel $\exp(-\lambda f(x, y))$ is positive definite for all $\lambda > 0$ if and only if $f$ is negative definite.

Proof: The original proof of this theorem can be found in $[24]$. An alternative proof can be found in Theorem 3.2.2 of $[23]$.

Theorem 5. Let $X$ be a nonempty set and let $f : (X \times X) \to \mathbb{R}$ be a positive definite kernel such that, for some $\rho > 0$, $|f(x, y)| \leq \rho$ for all $(x, y) \in X \times X$. Then, if $g(t) = \sum_{n=0}^{\infty} \alpha_n t^n$ for $|t| \leq \rho$ and $\alpha_n \geq 0$ for all $n$, the composite function $g \circ f$ is again positive definite.

Proof: We refer the reader to Corollary 3.1.14 of $[23]$ for a detailed proof of a more general version of this theorem. One can also verify this result by recalling that, when $f$ is a positive definite kernel so is $a f^n$ for all $n \in \mathbb{Z}_+$ and $a \geq 0$.

Now, let $x, y \in S^n$. Recall that the geodesic distance on $S^n$ is given by

$$d_g(x, y) = \arccos(\pi) = \frac{\pi}{2} - \arcsin(x^T y)$$

where $\arccos : [-1, 1] \to [0, \pi]$ and $\arcsin : [-1, 1] \to [-\pi/2, \pi/2]$ are the usual inverse trigonometric functions
defined with standard ranges. Note also that the Taylor series expansion of the arcsin function is given by

\[
\arcsin(t) = \sum_{n=0}^{\infty} \frac{(2n)!}{2^{2n}(n!)^2(2n+1)} t^{2n+1}, \quad |t| \leq 1. \tag{5}
\]

Next, let us define \( g : [-1, 1] \rightarrow \mathbb{R} \) as

\[
g(t) = \sum_{n=0}^{\infty} \frac{(2n)!}{2^{2n}(n!)^2(2n+1)} t^{2n+1}, \quad |t| \leq 1. \tag{6}
\]

It is well known, and can be proven directly using Definition 2, that the linear kernel \( x^T y \) is positive definite. Furthermore, \( |x^T y| \leq 1 \) since \( x, y \in S^n \). Now, by invoking Theorem 5 with \( f(x, y) = x^T y, \rho = 1 \) and \( g(t) \) defined as in (6), we conclude that \( \arcsin(x^T y) \) is positive definite. Then, using (4) and Lemma 3 it follows that \( d_g(x, y) \) is negative definite.

We can now complete the proof of Theorem 1. Since \( d_g \) is negative definite, from Theorem 4, \( k_g(x, y) := \exp(-d_g(x, y)/2\sigma^2) \) is positive definite for all \( \sigma \).

V. COMBINING MANIFOLD-VALUED DESCRIPTORS

We now address the object recognition problem with multiple manifold-valued descriptors (features). Intuitively, object classification can be improved when multiple descriptors are used instead of just one descriptor. However, as stated earlier, when the different descriptors lie on different manifolds, combining them optimally to maximize the classification accuracy is not straightforward.

In the following, we first discuss the general theory of the proposed method and then proceed to addressing the special case of combining HOG and Region Covariance descriptors.

A. MKL on Multiple Manifolds

In order to combine descriptors lying on different manifolds, we propose to first embed each manifold in a Hilbert space using a positive definite kernel defined on that manifold, and then to combine multiple features in those Hilbert spaces. In [14], it was shown that optimally combining multiple features lying of different Hilbert spaces can be done via Multiple Kernel Learning (MKL).

We now formalize our method for binary classification. It can easily be extended to the multi-class case using one of the standard one-vs-all and one-vs-one procedures.

Let \( \{(I_i, y_i)\}_{i=1}^{m} \) be a given set of training samples, where each \( I_i \in \mathcal{X} \) (the set of all possible images) is an image and \( y_i \in \{-1, 1\} \) is the label of image \( I_i \). Furthermore, let \( \{g_j\}_{j=1}^{n} \) be a set of descriptor generating functions where each \( g_j : \mathcal{X} \rightarrow \mathcal{M}_j \) generates a descriptor lying on the manifold \( \mathcal{M}_j \). We aim to learn a binary classifier \( f : \mathcal{X} \rightarrow \{-1, 1\} \) by selecting and optimally combining the different descriptors lying of the manifolds \( \mathcal{M}_1, \ldots, \mathcal{M}_N \).

We assume the existence of a positive definite kernel \( k_j \) on each \( \mathcal{M}_j \). Let \( K^{(j)} \) be the kernel matrix generated by \( g_j \) and \( K^{(j)}_{pq} = k_j(g_j(I_p), g_j(I_q)) \). The overall kernel is then expressed as \( K^* = \sum_j a_j K^{(j)} \), where each \( a_j \geq 0 \).

The MKL procedure described above has been shown to outperform other popular feature selection/combination methods such as wrappers, filters and boosting [25].

B. Combining HOG and Region Covariance Descriptors

Both HOG and Region Covariance descriptors have been immensely successful as region descriptors in object classification/detection. Intuitively, a harmony between these two descriptors should further enhance the object classification accuracy.

Given an image window \( I_i \in \mathcal{X} \), we use two descriptor generating functions \( g_1 : \mathcal{X} \rightarrow S^n \) and \( g_2 : \mathcal{X} \rightarrow S_{ym}^n \), the former generating HOG descriptors and the latter generating covariance descriptors. HOG descriptors are scaled to have unit norm after block-normalization. As was discussed earlier, this scaling does not result in any distortion of the descriptors. We use our geodesic exponential kernel \( k_{g_1} \) on \( S^n \) (Theorem 1) and the Log-Euclidean Gaussian kernel \( k_{LE} \) on \( S_{ym} \) (defined in (3)) as positive definite kernels on the two manifolds and employ the MKL procedure described in the previous section. For the multi-class case, we use the standard one-vs-all procedure. As will be shown in the experimental evaluation, classifiers learnt using both HOG and Region Covariance descriptors indeed significantly outperform the classifiers learnt using only one of the descriptors.

VI. EXPERIMENTS

In this section we provide an experimental evaluation of the proposed method. Details for our implementations and evaluation methodology for each dataset will be described in the following subsections.

A. ETH-80

The ETH-80 dataset [26] contains images of 8 different classes (apple, car, cow, cup, dog, horse, pear, and tomato), with 10 objects per class (eg. 10 different apples) and 41 images per object taken from different 3D angles. Sample images from the dataset are shown in Fig.1. In our experiment, we randomly picked 5 images per object for training and used the remaining 36 for testing. This random splitting process was repeated 5 times and we report the average accuracy. The goal was to predict the correct class of a given test image.

All ETH images have a common size of \( 256 \times 256 \). To calculate HOG descriptors, we used 16 \( \times \) 16-dimensional cells, 9 histogram bins per cell and 4-fold block normalization. The resulting HOG descriptor is a \( 15 \times 15 \times 9 \times 4 = 8100 \)-dimensional vector having a fixed norm of 225. We also used histogram clipping [3] where normalized histogram values are clipped at 0.2 and then re-normalized. The PMT toolbox [27] was utilized with some modifications to calculate the HOG.
B. Caltech-101 dataset

Our next experiment is on the Caltech-101 dataset [28]. This dataset contains images of 101 object classes with at least 30 images from each class. Object recognition on this dataset is very challenging due to the high number of classes and large intra-class variances. Sample images from the dataset are shown in Fig. 2. We followed the procedure suggested in the official dataset documentation [29], and randomly picked 15 training images and 15 test images per class. This random picking was repeated 5 times, and we report the average accuracy.

Since computation of HOG descriptors requires a common image size, we cropped objects using the annotations provided with the dataset and included them in $128 \times 128$ windows, centered, having resized as required. Both HOG and Region Covariance descriptors were calculated on these windows. We used the same procedure described in Section V.A, but with cells of dimension $8 \times 8$. The resulting HOG descriptors again have 8100 dimensions. Covariance descriptors were computed with the feature vector $[x \ y \ I \ |I_x| \ |I_y| \ |I_{xx}| \ |I_{yy}|]$ and therefore lie on $\text{Sym}^2_{+}$. The result-

Table 2 summarizes the classification accuracies for the different methods described in Section V.A. In addition to the methods described in that section, we also compared our method to other results previously reported on the same dataset with a similar evaluation methodology where only 15 training samples are used. Recognition accuracies for these methods were taken directly from the respective papers.

According to the results shown in Table II, we observe that covariance descriptors alone perform poorly on the Caltech-101 dataset. This is perhaps because we use only one covariance descriptor per image (calculated on the whole image). When used on challenging datasets, covariance descriptors are conventionally calculated on multiple patches [2]. [4] and then combined using a feature selection algorithm such as LogitBoost [4]. However, as can be seen from Table II even when computed only on the full window, covariance descriptors help boosting the HOG descriptors thanks to our method. It is also seen from the results that the proposed geodesic exponential kernel outperforms the Euclidean Gaussian kernel used on HOG descriptors. This might be attributed to the fact that our kernel accounts for the true geometry of $S^n$ while the Euclidean Gaussian kernel does not.

VII. Conclusion

We have proposed a feature combination/selection method to combine multiple features lying on different manifolds. As an application of the proposed method, we have tackled the problem of combining HOG descriptors and Region Covariance descriptors. To this end, we have proposed a new
kernel on the manifold where HOG descriptors lie, i.e., the unit sphere $S^n$, that accounts for the Riemannian geometry of the manifold. We have also made use of a previously proposed Riemannian kernel on covariance descriptors. Our experiments on two different datasets show that the proposed method outperforms using HOG and Region Covariance descriptors individually.

Although we only demonstrated combining multiple manifold-valued features in an SVM classifier, it is possible to use the same idea of multi-manifold kernel learning in a kernel Fisher discriminant analysis based classifier [32]. This could be an interesting area for future research.

**REFERENCES**


