Image Set Classification by Symmetric Positive Semi-Definite Matrices

Masoud Faraki Mehrtash T. Harandi Fatih Porikli
Research School of Engineering, Australian National University, Australia
NICTA, Australia
{masoud.faraki,mehrtash.harandi,fatih.porikli}@nicta.com.au

Abstract

Representing images and videos by covariance features and leveraging the inherent manifold structure of symmetric positive definite (SPD) matrices leads to enhanced performances in various visual recognition tasks. However, when covariance features are used to represent image-sets, the result is often rank-deficient. Thus, most existing approaches adhere to blind perturbation with predefined regularizers just to be able to employ inference tools.

To overcome this problem, we introduce novel similarity measures specifically designed for rank-deficient covariance features, i.e.*, symmetric positive semi-definite (SPSD) matrices. In particular, we derive positive definite kernels that can be decomposed into the kernels on the cone of SPD matrices and kernels on the Grassmannian manifold. Using the standard test protocols, our method achieves superior results for image set classification on YouTube Celebrities, Cambridge Hand Gesture, and Maryland Dynamic Scene benchmarks.*

1. Introduction

Symmetric positive semi-definite (SPSD) matrices naturally arise for applications where the number of observed samples is lower than the dimensionality of the samples, and a covariance matrix is used to represent the observations. One such application is image set classification where each set contains a number of images that belong to the same class. Compared to single image based classification, recognition from image sets has a significant advantage of efficiently exemplifying intra-class appearance variations such as pose changes, illumination differences, partial occlusions and object deformations through multiple representatives [7, 19]. Therefore, proper modeling of image sets permits utilizing intra-class variation in the set as a complementary cue, thus enables discriminative representations [26].

Covariance features provide rich yet compact representations for image set modeling as they allow fusing various image cues while attenuating the impact of noisy samples through their averaging process [36, 12]. Moreover, modern inference frameworks [11, 9, 5, 37, 10] are available for symmetric positive definite (SPD) matrices.

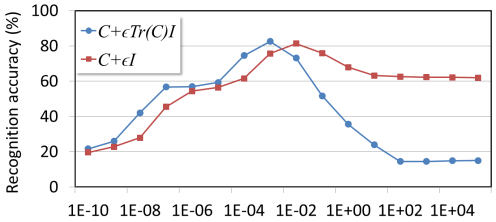

Figure 1: Recognition performance of a conventional NN classifier using full rank matrices by regularizing the rank-deficient covariance feature. As seen, the performance changes drastically from 15% to 82% for different values of the regularization parameter ϵ.

An SPSD matrix is a result of constructing a covariance feature for an image set by arranging the d-dimensional vector descriptors of p images into the columns of a tall matrix $\boldsymbol{X}$ of size $d \times p$. Since the dimensionality of the image descriptor is often several orders of magnitude greater than the number of images, *i.e.* $p \ll d$, the covariance feature constructed from the set is a rank-deficient SPSD matrix.

An important issue here is that, most of the existing tools developed upon the manifold of SPD matrices including its natural metric, distance measures, and statistics are only valid for full-rank matrices. Hence, previous studies [36, 17] adhere to ad-hoc solutions to overcome the rank deficiency by perturbing the rank-deficient covariance matrix $\boldsymbol{C}$ with a constant regularizer, *e.g.* $\boldsymbol{C} + \epsilon \mathrm{I}_d$ adding a scaled identity matrix I_d. Such a regularization nonetheless may deteriorate the performance as shown in Figure 1. This is a very practical, albeit overlooked, problem lacking of a competent solution.

In this paper, we overcome the above issue with novel methods by proposing a negative definite distance metric (see Sec.4.1) inspired by the indefinite *closeness* measure [5], and incorporating into positive definite kernels for image set classification task. The proposed negative definite metric enjoys several desirable properties (*e.g.*, invariance

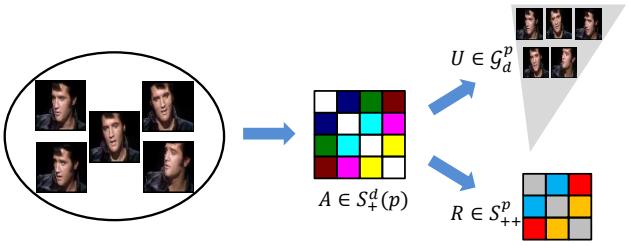

Figure 2: A conceptual example of our proposed image set representation. Image set is represented by an SPSD matrix A which is further decomposed to a linear subspace U and an SPD matrix R.

to rotation) and is constituted of two parts, a linear subspace and a smaller SPD matrix (see Figure 2 for a conceptual illustration).

We then turn our attention to derive positive definite kernels including linear, polynomial, Laplace, and RBF, based on the negative definite metric. For this, we make embed the curved product space of $\mathcal{G}_d^p$ and $\mathcal{S}_{++}^p$ to the space of symmetric matrices obtained via the projection distance [15] and the log-Euclidean distance [3], respectively. Subsequently, as a classifier we use kernel Discriminant Analysis (kDA) [27] that employs the kernel trick to perform linear discriminant analysis in a high-dimensional feature space in order to extract the significant nonlinear features which maximize the between-class variance and minimize the within-class variance. In other words, we generalize discriminative power of kDA to the manifold of SPSD matrices.

Our experiments demonstrate the superiority of the proposed methods against several baseline and state-of-the-art methods. To the best of our knowledge, using the standard testing protocol, our method with the proposed kernels obtained in this new geometry equipped with a kDA classifier achieves the best reported results on standard image set classification benchmarks: 72.8% for YouTube celebrities face recognition [22], 91.1% for Cambridge hand gesture recognition [24], and 90.0% for Maryland dynamic scene recognition [33].

2. Related Work

Almost all image set classification techniques have to make two major decisions: (i) how to represent an image set, and (ii) what metric to use to measure the similarity between sets.

From the representation point of view, existing solutions can be divided roughly into model-driven and topology-driven approaches. As for the model-driven methods such as [25, 28]), it is usually assumed that the images within a set are the samples from a known parametric form. The notable examples include modeling sets by single Gaussian distribution [32] and Gaussian Mixture Models (GMM) [2]. Once the model for each image-set is obtained, the similarity between sets can be obtained either as the distance between models (*e.g.*, Kullback-Leibler (KL) divergence between Gaussian models) or more directly as the distance between the estimated parameters. The performance of model-driven methods will deteriorate if the set data is weakly correlated to the model.

To alleviate this difficulty, the topology-driven methods assume data establish a topological space and represent image sets by sophisticated nonlinear manifolds [23, 38, 15, 18, 37, 7]. Kim *et al*. in [23] learn a discriminant function that maximizes the canonical correlations of within-class sets while minimizing the canonical correlations of between-class sets. Then, image sets transformed by the discriminant function are compared by the canonical correlations with transformed subspaces. The sum of the cosines of the principal angles have been successfully utilized in [38] for image sets represented by linear subspaces. Harandi *et al*. [18] propose a discriminant analysis approach on Grassmannian manifolds, based on a graph embedding framework. They show that by introducing within-class and between-class similarity graphs to characterize intra-class compactness and inter-class separability, the correct geometrical structure of data can be exploited. Wang *et al*. in [37] model image sets by their natural second-order statistics, *i.e.*, covariance matrices. Since nonsingular covariance matrices lie on a Riemannian manifold, a kernel function is used to explicitly embed the Riemannian structure into a Euclidean space. Chen *et al*. [7] achieved improved performances by computing the distance between different locally linear subspaces. By taking the advantage of the underlying geometrical structure, topology-driven methods provide robustness to noise and can operate with a relatively small number of samples per class.

Affine hull approaches [20, 6], on the other hand, adaptively choose optimal samples to obtain geometric distances between image sets instead of considering the structure of all data samples. As a consequence, they allow larger intra-class variation, which results to a more general image set modeling for challenging applications such as face recognition in the wild. Nevertheless, misclassification can occur if the nearest points between two hulls are not correctly labeled.

In line with the methods that represent image sets on some geometric surfaces, very recently Hayat *et al*. [19] learn class-specific models by an Adaptive Deep Network Template (ADNT). Based on the minimum reconstruction error from the learned models, a majority voting strategy is used for classification.

In practice, restricting the model or topology to obey some form of predefined structure *e.g*. linear subspaces, statistical distributions, and etc. will result in loss of generality

and degraded performance.

3. Preliminaries

We compare against nearest-neighbor classifiers using geodesic distances in the corresponding Riemannian manifolds. Moreover, the closeness measure and our negative definite distance metrics are derived upon the Riemannian geometry. Hence, we briefly summarise basic Riemannian concepts for completeness here.

Riemannian Manifold: A manifold $\mathcal{M}$ is a topological space which is locally homeomorphic to the $d-$dimensional Euclidean space $\mathbb{R}^d$, for some d called the dimensionality of the manifold. The tangent space attached to a point $\boldsymbol{X}$ on the manifold, $T_{\boldsymbol{X}}\mathcal{M}$, is a vector space that consists of the tangent vectors of all possible curves passing through $\boldsymbol{X}$. A Riemannian manifold is a differential manifold with a metric defined on the tangent spaces.

SPD Manifold: The space of $d \times d$ SPD matrices endowed with a Riemannian metric forms a Riemannian manifold [30]. The affine invariant Riemannian metric (AIRM) is the most popular choice to handle the non-Euclidean structure of SPD matrices and is shown to be advantageous for several applications [35]. For $\boldsymbol{X} \in \mathcal{S}_{++}^d$ and two tangent vectors $\Delta_1, \Delta_2 \in T_{\boldsymbol{X}}\mathcal{M}$, the AIRM is defined as

$$\begin{aligned}\langle \Delta_1, \Delta_2 \rangle_{\boldsymbol{X}} &\triangleq \langle \boldsymbol{X}^{-1/2}\Delta_1\boldsymbol{X}^{-1/2}, \boldsymbol{X}^{-1/2}\Delta_2\boldsymbol{X}^{-1/2} \rangle \\ &= \mathrm{Tr}\left(\boldsymbol{X}^{-1}\Delta_1\boldsymbol{X}^{-1}\Delta_2\right) \ . \end{aligned} \tag{1}$$

For two $\boldsymbol{X}, \boldsymbol{Y} \in \mathcal{S}_{++}^d$, the geodesic distance induced by AIRM is

$$\delta_g(\boldsymbol{X}, \boldsymbol{Y}) = \| \log(\boldsymbol{X}^{-1/2}\boldsymbol{Y}\boldsymbol{X}^{-1/2}) \|_F \ , \tag{2}$$

where $\|.\|_F$ denotes the Frobenius norm and $\log(\cdot)$ is the matrix principal logarithm.

In addition to AIRM, the log-Euclidean metric [3] $\delta_L : \mathcal{S}_{++}^d \times \mathcal{S}_{++}^d \rightarrow [0, \infty)$ is widely used to measure similarities on SPD manifolds. It is defined as

$$\delta_L(\boldsymbol{X}, \boldsymbol{Y}) \triangleq \| \log(\boldsymbol{X}) - \log(\boldsymbol{Y}) \|_F \ , \tag{3}$$

where $\|.\|_F$ denotes the Frobenius norm and $\log(\cdot)$ is the matrix principal logarithm.

Grassmannian Manifold: The space of p-dimensional linear subspaces of $\mathbb{R}^d$ for $0 < p < d$ is a Riemannian manifold known as the Grassmannian $\mathcal{G}_d^p$ [1]. A point on the Grassmannian manifold $\mathcal{G}_d^p$ might be specified by an arbitrary $d \times p$ matrix with orthogonal columns, *i.e.*, $\boldsymbol{X} \in \mathcal{G}_d^p \Rightarrow \boldsymbol{X}^T\boldsymbol{X} = \mathbf{I}_p$. Such a point corresponds to a subspace spanned by the columns of a $d \times p$ full rank matrix and therefore is denoted by $\mathrm{span}(\boldsymbol{X})$.

For two tangents Δ_1 and Δ_2 at $\boldsymbol{X}$ the Riemannian metric is defined as follows

$$\langle \Delta_1, \Delta_2 \rangle_{\boldsymbol{X}} = \mathrm{Tr}\left({\Delta_1}^T\Delta_2\right) \ . \tag{4}$$

Using this, the geodesic distance between two points $\boldsymbol{X}$ and $\boldsymbol{Y}$ is given by

$$\delta_g(\boldsymbol{X}, \boldsymbol{Y}) = \|\Theta\|_2 \ , \tag{5}$$

where Θ is the vector of principal angles between $\boldsymbol{X}, \boldsymbol{Y}$.

In addition to the geodesic distance, another popular distance in $\mathcal{G}_d^p$ is the projection distance $\delta_P : \mathcal{G}_d^p \times \mathcal{G}_d^p \rightarrow \mathbb{R}^+$ [16, 15] defined as

$$\delta_P^2(\boldsymbol{X}, \boldsymbol{Y}) = \|\boldsymbol{X}\boldsymbol{X}^T - \boldsymbol{Y}\boldsymbol{Y}^T\|_F^2, \tag{6}$$

where $\| \cdot \|_F$ denotes the Frobenius norm. The projection metric is related to the geometry of Grassmannian developed in [8]. Moreover, the length of any curve is the same under δ_p and δ_g up to a scale of $\sqrt{2}$.

4. Riemannian Metric for SPSD Matrices

As mentioned earlier, computing the covariance features from vectorized image features (*i.e.* raw intensity or any other image descriptors such as histograms) results in rank-deficient matrices due to the fact that the dimensionality of the features is greater than the number of images in the set. As a result, the covariance features become an instance of SPSD matrices. We utilize the Riemannian metric for SPSD matrices of fixed-rank introduced in [5] to handle such covariance features. The metric addresses weaknesses of the natural metric in SPD manifold in dealing with rank-deficient matrices while enjoying several invariance properties. More specifically, the metric leads to a natural metric with decoupled contributions in Grassmannian and SPD manifolds.

Let $\boldsymbol{I} = \left[\vec{I}_1 | \vec{I}_2 | \cdots | \vec{I}_p\right], \vec{I}_i \in \mathbb{R}^d$ be a $d \times p$ matrix of p observations. Then, the covariance feature $\boldsymbol{C}$ is formally defined as

$$\boldsymbol{C} = \frac{1}{p-1}\sum_{i=1}^{p}(\vec{I}_i - \boldsymbol{\mu})(\vec{I}_i - \boldsymbol{\mu})^T, \tag{7}$$

where $\boldsymbol{\mu} = \frac{1}{p}\sum_{i=1}^{p}\vec{I}_i$ is the sample mean of the observations.

When $d > p$, $\boldsymbol{C}$ is rank-deficient, which means that the resulting matrix would be on the boundary of the positive cone. As a result, one might totally dismiss the luxury of computational tools in SPD manifold to analyse such covariance features. For instance, the distance from any SPD matrix to $\boldsymbol{C}$ would be infinite according to the AIRM. To overcome this issue, off-the-shelf treatment (for example proposed in [37]) is through regularizing the original $\boldsymbol{C}$, *i.e.*,

$$\boldsymbol{C}^* = \boldsymbol{C} + \epsilon \mathrm{I}_d \ , \tag{8}$$

where ϵ is a constant and I_d is the $d \times d$ identity matrix.

As we will show in our experiments, the perturbation deteriorates the discriminatory power of covariance features.

Here, we are interested in taking the advantage of true geometry of the resulting covariance features. We commence by deriving the natural metric and the geodesic distance for SPSD matrices of fixed-rank and then turn our attention to create valid kernels.

From quotient manifold perspective, points on $\mathcal{G}_d^p$ are yielded by grouping points on Steifel manifold $\mathcal{S}_d^p$, the set of $d \times p$ matrices with orthogonality constraint, that represent the same subspace [1]. Therefore, $\mathcal{G}_d^p$ admits the following quotient manifold representation

$$\mathcal{G}_d^p \cong \mathcal{S}_d^p / \mathcal{O}_p \ , \tag{9}$$

where $\mathcal{O}_p$ denotes the orthogonal group in dimension p.

Let $\boldsymbol{A} \in \mathcal{S}_+^d(p)$, obtained for example from computing the empirical covariance matrix of an image set. For any such matrix, there exists the following decomposition

$$\boldsymbol{A} = \boldsymbol{Z}\boldsymbol{Z}^T = (\boldsymbol{U}\boldsymbol{R})(\boldsymbol{U}\boldsymbol{R})^T = \boldsymbol{U}\boldsymbol{R}^2\boldsymbol{U}^T \ , \tag{10}$$

where $\boldsymbol{Z}$ is a full-rank $d \times p$ matrix, $\boldsymbol{U} \in \mathcal{S}_d^p$, and $\boldsymbol{R}^2 \in \mathcal{S}_{++}^p$.

Eqn (10) remains unchanged under the transformation $\boldsymbol{Z} \rightarrow \boldsymbol{Z}\boldsymbol{O}$ for any matrix $\boldsymbol{O} \in \mathcal{O}_p$. Thus, one can deduce that the equivalence relation $(\boldsymbol{U}, \boldsymbol{R}^2) \equiv (\boldsymbol{U}\boldsymbol{O}, \boldsymbol{O}^T\boldsymbol{R}^2\boldsymbol{O})$ holds. As a result, the set $\mathcal{S}_+^d(p)$ admits the the quotient manifold representation $\mathcal{S}_+^d(p) \cong \left(\mathcal{S}_d^p \times \mathcal{S}_{++}^p\right)/\mathcal{O}_p$.

The metric proposed by [5] is defined to be the sum of infinitesimal distances in $\mathcal{G}_d^p$ and $\mathcal{S}_{++}^p$. Let $\triangle$ and $\boldsymbol{D}$ represent the tangent vectors in Grassmannian and SPD manifolds, respectively. For $\mathcal{S}_+^d(p) \ni \boldsymbol{A} = \boldsymbol{U}\boldsymbol{R}^2\boldsymbol{U}^T$ and two pair of tangent vectors $(\triangle_{\mathbf{1}}, \boldsymbol{D}_1)$ and $(\triangle_{\mathbf{2}}, \boldsymbol{D}_2)$ the metric is defined as

$$\begin{aligned}\langle(\triangle_{\mathbf{1}}, \boldsymbol{D}_1),(\triangle_{\mathbf{2}}, \boldsymbol{D}_2)\rangle_{\boldsymbol{A}} := \\ \langle\triangle_{\mathbf{1}}, \triangle_{\mathbf{2}}\rangle + \lambda\langle\boldsymbol{R}^{-1}\boldsymbol{D}_1\boldsymbol{R}^{-1}, \boldsymbol{R}^{-1}\boldsymbol{D}_2\boldsymbol{R}^{-1}\rangle \ ,\end{aligned} \tag{11}$$

where $\langle\cdot,\cdot\rangle$ denotes the normal inner product and $\lambda \geq 0$ is the combination weight.

Following Eqn (10), for two SPSD matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathcal{S}_+^d(p)$ we obtain $\boldsymbol{A} = \boldsymbol{U}_A\boldsymbol{R}_A^2\boldsymbol{U}_A^T$ and $\boldsymbol{B} = \boldsymbol{U}_B\boldsymbol{R}_B^2\boldsymbol{U}_B^T$. Then, the metric induces the following (squared) geodesic distance between $\boldsymbol{A}$ and $\boldsymbol{B}$

$$\delta_g^2(\boldsymbol{A}, \boldsymbol{B}) = \|\Theta\|_F^2 + \lambda\|\log(\boldsymbol{R}_{\boldsymbol{A}}^{-1}\boldsymbol{R}_{\boldsymbol{B}}^2\boldsymbol{R}_{\boldsymbol{A}}^{-1})\|_F^2 \ , \tag{12}$$

with $\lambda \geq 0$. The chosen metric is simply the sum of infinitesimal distances in $\mathcal{G}_d^p$ and $\mathcal{S}_{++}^p$. The first term refers to the squared geodesic distance between linear subspaces $\boldsymbol{U}_A$ and $\boldsymbol{U}_B$ while the second term is the squared geodesic distance between two SPD matrices $\boldsymbol{R}_A^2$ and $\boldsymbol{R}_B^2$. Moreover, the distance is invariant to angle preserving transformations (*i.e.* orthogonal transformations, scalings, and pseudoinversion). Here, our main motivation to benefit from the manifold of SPSD matrices is to overcome the limitations of the SPD manifolds in dealing with rank deficient matrices.As will be demonstrated by our experiments, the induced geometry is more discriminative than both SPD and Grassmannian manifolds.

4.1. Kernels on SPSD Matrices

To define positive definite (*pd*) kernels on the SPSD manifold, we first obtain a negative definite (*nd*) function on $\mathcal{S}_+^d(p)$.

Theorem 1. *The function* $\delta^2 : \mathcal{S}_+^d(p) \times \mathcal{S}_+^d(p) \rightarrow \mathbb{R}_+$ *defined as*

$$\begin{aligned}\delta^2(\boldsymbol{A}, \boldsymbol{B}) &\triangleq \|\boldsymbol{U}_A\boldsymbol{U}_A^T - \boldsymbol{U}_B\boldsymbol{U}_B^T\|_F^2 + \lambda\|\log(\boldsymbol{R}_A) - \log(\boldsymbol{R}_B)\|_F^2 \\ &= 2p - 2\|\boldsymbol{U}_A^T\boldsymbol{U}_B\|_F^2 + \lambda\|\log(\boldsymbol{R}_A) - \log(\boldsymbol{R}_B)\|_F^2 \ ,\end{aligned} \tag{13}$$

is negative definite on $\mathcal{S}_+^d(p)$ *for* $\lambda \geq 0$.

Proof. We recall that a symmetric function $\psi : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ on a set $\mathcal{X}$ is *nd* if and only if $\sum_{i,j=1}^n c_i c_j k(x_i, x_j) \leq 0$ for any $n \in \mathbb{N}$, $x_i \in \mathcal{X}$ and $c_i \in \mathbb{R}$ with $\sum_{i=1}^n c_i = 0$. As shown in [21], if $f : \mathcal{X} \rightarrow \mathcal{H}$ is a mapping from a set $\mathcal{X}$ to an inner product space $\mathcal{H}$, then the function $\|f(x_i) - f(x_j)\|_{\mathcal{H}}^2$ is negative definite for $\forall x_i, x_j \in \mathcal{X}$. Here $\|\cdot\|_{\mathcal{H}}$ denotes the norm in $\mathcal{H}$.

Now we note that $\pi_p : \mathcal{G}_d^p \rightarrow \mathrm{Sym}(d), \pi_p(\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{X}^T$ is a mapping from the Grassmannian to the space of $d \times d$ symmetric matrices, hence the first term on RHS of Eqn. (13). Similarly, with $\log : \mathcal{S}_{++}^p \rightarrow \mathrm{Sym}(p)$, the second term in the RHS of Eqn. (13) is negative definite. By invoking the definition of the negative definite kernels, it is easy to see that the addition of two negative definite kernels is also a negative definite kernel. □

Having a *nd* function at our disposal, we can make use of the following theorem to define a family of *pd* kernels on $\mathcal{S}_+^d(p)$.

Theorem 2 (Theorem 2.3 in Chapter 3 of [4])**.** *Let* μ *be a probability measure on the half line* $\mathbb{R}_+$ *and* $0 < \int_0^\infty t\mathrm{d}\mu(t) < \infty$. *Let* $\mathcal{L}_\mu$ *be the Laplace transform of* μ, i.e., $\mathcal{L}_\mu(s) = \int_0^\infty e^{-ts}\mathrm{d}\mu(t), \ s \in \mathbb{C}_+$. *Then,* $\mathcal{L}_\mu(\beta f)$ *is positive definite for all* $\beta > 0$ *if and only if* $f : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_+$ *is negative definite.*

For example, by choosing μ to be the Dirac function at $t = 1$, we obtain the RBF kernel on $\mathcal{S}_+^d(p)$ as follows

$$\begin{aligned}k_R(\boldsymbol{A}, \boldsymbol{B}) \triangleq \\ \exp\Big(-\beta\Big(\lambda\|\log(\boldsymbol{R}_A) - \log(\boldsymbol{R}_B)\|_F^2 - 2\|\boldsymbol{U}_A^T\boldsymbol{U}_B\|_F^2\Big)\Big) \ .\end{aligned} \tag{14}$$

We notice that one could arrive to the same conclusion, *i.e.*, $k_R(\cdot,\cdot)$ is *pd*, by observing that it is indeed the product of two *pd* kernels. However, our approach here is more principled and can be used to generate other types of *pd*

Figure 3: Examples of the YouTube celebrities dataset [22].

kernels on $\mathcal{S}_+^d(p)$ by properly changing the measure μ in Thm.2.

Another widely used kernel in the Euclidean spaces is the Laplace kernel defined as $k(\boldsymbol{x},\boldsymbol{y}) = \exp(-\beta\|\boldsymbol{x}-\boldsymbol{y}\|)$. To obtain the Laplace kernel on the $\mathcal{S}_+^d(p)$, we make use of the following theorem for *nd* kernels.

Theorem 3 (Corollary 2.10 in Chapter 3 of [4])**.** *If $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is negative definite and satisfies $f(\boldsymbol{x},\boldsymbol{x}) \geqq 0$ then so is ψ^α for $0 < \alpha < 1$*

As a result both $\delta(\cdot,\cdot) = \sqrt{\delta^2(\cdot,\cdot)}$ is *nd* by choosing $\alpha = 1/2$ in Theorem 3 and hence the form of $\exp(-\beta\delta(\cdot,\cdot))$ is *pd*.

Before concluding this part, we also introduce the linear and polynomial kernels on $\mathcal{S}_+^d(p)$. The linear kernel $k_l(\boldsymbol{A},\boldsymbol{B}) = \|\boldsymbol{U}_A^T\boldsymbol{U}_B\|_F^2 + \lambda \operatorname{Tr}\left(\log(\boldsymbol{R}_A)\log(\boldsymbol{R}_B)\right)$ is interesting as it is a parameter-less kernel (discarding λ which defines the form of the linear combination of the two). To show that $k_l(\cdot,\cdot)$ is *pd*, we note that $k_l(\cdot,\cdot)$ is the summation of two *pd* kernels defined on the space of symmetric matrices. This will lead us to define the polynomial kernels as

$$k_p(\boldsymbol{A},\boldsymbol{B}) \triangleq \left(\beta + \|\boldsymbol{U}_A^T\boldsymbol{U}_B\|_F^2 + \lambda \operatorname{Tr}\left(\log(\boldsymbol{R}_A)\log(\boldsymbol{R}_B)\right)\right)^\alpha. \quad (15)$$

We show all kernels we introduced in Table 1.

5. Experiments

We present experiments on three benchmark image set classification tasks. In all our experiments, we used a feature which suits the application. Notice that, our goal in this work is not feature selection.

We utilized the log-Euclidean and the projection distance. We tested different classifiers: and relied on two different classifiers: a simple nearest neighbor (NN) classifier to crystallize the benefits of using the proposed formulation and a KDA based classifier with the positive definite kernels we introduced in this paper. Different algorithms tested in our experiments are referred to as

NN: Nearest Neighbour classifier using the geodesic distance.

KDA$_{Linear}$: KDA classifier with linear kernel.

KDA$_{Polynomial}$: KDA classifier with polynomial kernel.

KDA$_{Laplace}$: KDA classifier with Laplace kernel.

KDA$_{RBF}$: KDA classifier with RBF kernel.

5.1. Video-Based Face Recognition

In our first experiment, we tackled the task of video-based face recognition. To this end, we considered the YouTube celebrity dataset [22] which contains 1910 videos of 47 people (see Fig. 3 for a few examples). The large diversity of poses, illumination, and facial expressions in addition to high compression ratio of face images have made it the most challenging dataset for image set classification based face recognition.

For our evaluation, we followed the standard five-fold cross validation protocol used in [20, 37, 19] which divides the whole dataset equally (with minimum overlap) into five folds with 9 videos per subject in each fold. Three of the videos were randomly selected for training, while the remaining six were used for testing. We generated linear subspaces of order 6 by grouping features of individual frames.

From each video, we extracted the face regions using the tracker of Ross *et al*. [31]. We considered Local Binary Patterns (LBP) [29] as our feature. Each face region was divided into 2×2 distinct non-overlapping blocks and the features were extracted for each patch and concatenated to form the final frame descriptors. Therefore, each descriptor belongs to $\mathcal{S}_{++}^6$ and $\mathcal{G}_{232}^6$ for the covariance features and linear subspaces.

Table 2 summarizes the average recognition rates of all the studied methods. Several conclusion can be drawn here. First of all, we note that in all cases the new SPSD manifold achieves descent accuracy scores.Furthermore, a single RBF kernel in the SPSD manifold comfortably outperform all the state-of-the-art algorithms. We achieve average accuracy score of 72.8% which outperforms the closest competitor by 1.4% percentage points.

5.2. Hand Gesture Recognition

We performed another experiment to classify image sequences of hand gestures. To this end, we used the Cambridge hand gesture dataset [24] which contains 900 image sets of 9 gesture classes with large intra-class variations. The gestures are defined by 3 primitive hand shapes and 3 primitive motions (see Fig. 4 for examples). Therefore, the target task for this data set is to classify different shapes as well as different motions at a time.

We followed the experimental protocol suggested by Mahmood *et al*. [26] in which 100 image sets of each class

Table 1: The proposed SPSD kernels.

Kernel	Equation
Linear	$k_l(\boldsymbol{A},\boldsymbol{B}) = \|\boldsymbol{U}_A^T\boldsymbol{U}_B\|_F^2 + \lambda \operatorname{Tr}\left(\log(\boldsymbol{R}_A)\log(\boldsymbol{R}_B)\right)$
Polynomial	$k_p(\boldsymbol{A},\boldsymbol{B}) = \left(\beta + \|\boldsymbol{U}_A^T\boldsymbol{U}_B\|_F^2 + \lambda \operatorname{Tr}\left(\log(\boldsymbol{R}_A)\log(\boldsymbol{R}_B)\right)\right)^\alpha$
Laplace	$k_L(\boldsymbol{A},\boldsymbol{B}) = \exp\left(-\beta\sqrt{\lambda\|\log(\boldsymbol{R}_A)-\log(\boldsymbol{R}_B)\|_F^2 - 2\|\boldsymbol{U}_A^T\boldsymbol{U}_B\|_F^2}\right)$
RBF	$k_R(\boldsymbol{A},\boldsymbol{B}) = \exp\left(-\beta\left(\lambda\|\log(\boldsymbol{R}_A)-\log(\boldsymbol{R}_B)\|_F^2 - 2\|\boldsymbol{U}_A^T\boldsymbol{U}_B\|_F^2\right)\right)$

Table 2: Recognition scores for the YouTube celebrities [22].

SANP	65.0 [20]
CDL	70.1 [36]
ADNT	71.4 [19]
NN	65.3
KDA$_{RBF}$	**72.8**
KDA$_{Laplace}$	71.8
KDA$_{Polynomial}$	70.6
KDA$_{Linear}$	70.4

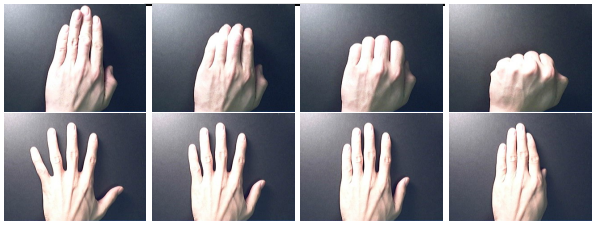

Figure 4: Examples of Cambridge hand gesture dataset [24].

Figure 5: Examples of the Maryland dynamic scene dataset [33].

Table 3: Recognition scores for the Cambridge hand gesture dataset [24].

SANP	22.5 [20]
CDL	73.4 [36]
SSSC	83.1 [26]
NN	87.4
KDA$_{RBF}$	**91.1**
KDA$_{Laplace}$	89.3
KDA$_{Polynomial}$	90.0
KDA$_{Linear}$	90.0

are divided into two parts, 81-100 used as train set and 1-80 as test set. For this dataset we made use of concatenated HOG features of 2×2 blocks of each frame. The state-of-the-art on this dataset [26] obtains the accuracy score of 83.1% using an ensemble of 9 spectral classifiers.

Table 3 shows that all the proposed methods comfortably outperform the state-of-the-art algorithms. A KDA classifier when the kernel is RBF over the SPSD manifold significantly outperforms the state-of-the-art ensemble of classifiers [26]. The difference is 8 percentage points.

5.3. Dynamic Scene Recognition

Finally, we considered the task of scene recognition from the videos using the Maryland "In-The-Wild"dataset [33] (see Fig. [24] for example classes). This dataset consists of 130 videos of natural scenes spanning 13 categories (*e.g.* Avalanche, Forest Fire, Waves) with 10 videos per class. The videos are collected from Internet-based video hosting sites, such as YouTube. Significant camera motions, differences in appearance, frame rate, scale, viewpoint, scene cuts, and illumination conditions exist in this dataset. A leave-one-video-out experimental protocol is used for consistency with previous evaluation in [14].

We made use of the FC7 features of Convolutional Neural Network (CNN) of Zhou *et al.* [39]. The network is trained on the Places dataset [39] with 205 scene categories and 2,5 millions of images with a category label. Here, we extract the 4096 FC7 feature of each frame. We then reduce the dimension of the feature to 400 using Principal Component Analysis.

Results are reported in Table 4. The table is self explanatory. To the best of our knowledge, 77.7% classification accuracy by the recent Bag of Spatiotemporal Energy (BoSE) method of Feichtenhofer *et al.* [14] is the highest accuracy score reported on this dataset. Our methods outperform the BoSE by a very large support.

5.4. Sensitivity to Rank and Weighting Parameter

We also studied the sensitivity of our proposed approach to the chosen subspace order as well as the value of λ. Figure 6 shows the accuracy against subspace order for the Cambridge hand gesture dataset using the pixel intensities as features and NN as classifier (*i.e.*, using Eqn 12). As depicted in the figure for all the studied subspace order the accuracy of NN on the Grassmannian manifold is inferior to the SPSD cases.

More importantly, we observed that as the order of the

Table 4: Recognition accuracies for the Maryland dataset [33].

SFA	60.0 [34]
CSO	67.7 [13]
BoSE	77.7 [14]
NN	83.1
KDA$_{RBF}$	88.5
KDA$_{Laplace}$	82.3
KDA$_{Polynomial}$	**90.0**
KDA$_{Linear}$	89.2

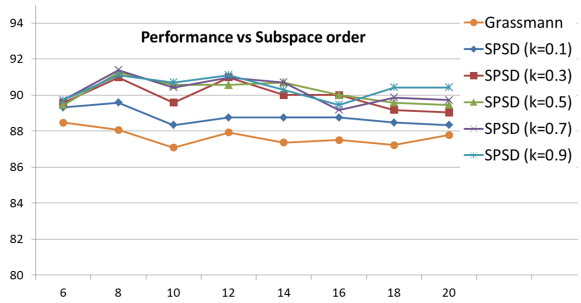

Figure 6: Accuracy against subspace order for the Cambridge hand gesture dataset. As visible inclusion of the SPD term significantly improves upon the use of Grassmannian only.

subspaces increases the differences between the accuracy obtained on the Grassmannian drops significantly. In other words, most values of the parameter λ provides a consistently stable performance over a range of p values even if the number of subspaces varies considerably. This clearly justifies the use of SPSD matrices.

6. Conclusions

Inspired by the recent success of image set representation as points on nonlinear Riemannian manifolds, we proposed Symmetric Positive Semi Definite (SPSD) matrices as descriptors for image set classification. The challenge lies in the fact that to measure the similarities, the usual metrics on the manifold of Symmetric Positive Definite (SPD) matrices, such as the Affine Invariant Riemannian Metric (AIRM), are not valid due to rank-deficiency of the SPSD matrices. Hence, our main motivation to benefit from the SPSD matrices is to overcome the limitations of the SPD manifolds (rank deficiency being the most important one).

We made use of a metric than can be decomposed as sum of infinitesimal distances on the Grassmannian manifold and the manifold of SPD matrices. Since our formulation enables us to utilize any distances on the two manifolds, we can integrate valid kernels for the image set classification task. A rigorous set of successful experiments on several challenging applications including video-based face recognition, gesture classification, and dynamic scene recognition demonstrated the advantages of our method.

Acknowledgements

NICTA is funded by the Australian Government through the Department of Communications, as well as the Australian Research Council through the ICT Centre of Excellence program.

References

[1] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, USA, 2008. 3, 4

[2] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell. Face recognition with image sets using manifold density divergence. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 581–588. IEEE, 2005. 2

[3] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications*, 29(1):328–347, 2007. 2, 3

[4] C. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups*. Springer, 1984. 4, 5

[5] S. Bonnabel and R. Sepulchre. Riemannian metric and geometric mean for positive semidefinite matrices of fixed rank. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1055–1070, 2009. 1, 3, 4

[6] H. Cevikalp and B. Triggs. Face recognition based on image sets. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2567–2573. IEEE, 2010. 2

[7] S. Chen, C. Sanderson, M. T. Harandi, and B. C. Lovell. Improved image set classification via joint sparse approximated nearest subspaces. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 452–459. IEEE, 2013. 1, 2

[8] A. Edelman, T. A. Arias, and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998. 3

[9] M. Faraki, M. T. Harandi, and F. Porikli. Approximate infinite-dimensional region covariance descriptors for image classification. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 1364–1368. IEEE, 2015. 1

[10] M. Faraki, M. T. Harandi, and F. Porikli. Material classification on symmetric positive definite manifolds. In *Proc. IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 749–756, 2015. 1

[11] M. Faraki, M. T. Harandi, and F. Porikli. More about vlad: A leap from euclidean to riemannian manifolds. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4951–4960. IEEE, 2015. 1

[12] M. Faraki, M. Palhang, and C. Sanderson. Log-euclidean bag of words for human action recognition. *IET Computer Vision*, 9(3):331–339, 2014. 1

[13] C. Feichtenhofer, A. Pinz, and R. P. Wildes. Spacetime forests with complementary features for dynamic scene recognition. In *Proc. British Machine Vision Conference (BMVC)*, 2013. 7

[14] C. Feichtenhofer, A. Pinz, and R. P. Wildes. Bags of spacetime energies for dynamic scene recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2014. 6, 7

[15] J. Hamm and D. D. Lee. Grassmann discriminant analysis: a unifying view on subspace-based learning. In *Proc. Int. Conference on Machine Learning (ICML)*, pages 376–383. ACM, 2008. 2, 3

[16] M. Harandi, C. Sanderson, C. Shen, and B. Lovell. Dictionary learning and sparse coding on Grassmann manifolds: An extrinsic solution. In *Proc. Int. Conference on Computer Vision (ICCV)*, pages 3120–3127. IEEE, 2013. 3

[17] M. T. Harandi, M. Salzmann, and R. Hartley. From manifold to manifold: geometry-aware dimensionality reduction for spd matrices. In *Proc. European Conference on Computer Vision (ECCV)*, pages 17–32. Springer, 2014. 1

[18] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2705–2712. IEEE, 2011. 2

[19] M. Hayat, M. Bennamoun, and S. An. Learning non-linear reconstruction models for image set classification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1915–1922. IEEE, 2014. 1, 2, 5, 6

[20] Y. Hu, A. S. Mian, and R. Owens. Sparse approximated nearest points for image set classification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 121–128. IEEE, 2011. 2, 5, 6

[21] S. Jayasumana, R. Hartley, M. Salzmann, H. Li, and M. Harandi. Kernel methods on riemannian manifolds with gaussian rbf kernels. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 37(12):2464–2477, 2015. 4

[22] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley. Face tracking and recognition with visual constraints in real-world videos. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2008. 2, 5, 6

[23] T.-K. Kim, J. Kittler, and R. Cipolla. Discriminative learning and recognition of image set classes using canonical correlations. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 29(6):1005–1018, 2007. 2

[24] T.-K. Kim, K.-Y. K. Wong, and R. Cipolla. Tensor canonical correlation analysis for action classification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2007. 2, 5, 6

[25] X. Li, K. Fukui, and N. Zheng. Boosting constrained mutual subspace method for robust image-set based object recognition. In *Proc. Int. Joint Conference on Artificial Intelligence (IJCAI)*, pages 1132–1137, 2009. 2

[26] A. Mahmood, A. Mian, and R. Owens. Semi-supervised spectral clustering for image set classification. In *CVPR*, pages 121–128. IEEE, 2014. 1, 5, 6

[27] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller. Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing*, pages 41–48. IEEE, 1999. 2

[28] M. Nishiyama, M. Yuasa, T. Shibata, T. Wakasugi, T. Kawahara, and O. Yamaguchi. Recognizing faces of moving people by hierarchical image-set matching. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2007. 2

[29] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 24(7):971–987, 2002. 5

[30] X. Pennec, P. Fillard, and N. Ayache. A riemannian framework for tensor computing. *Int. Journal of Computer Vision (IJCV)*, 66(1):41–66, 2006. 3

[31] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang. Incremental learning for robust visual tracking. *Int. Journal of Computer Vision (IJCV)*, 77(1-3):125–141, 2008. 5

[32] G. Shakhnarovich, J. W. Fisher, and T. Darrell. Face recognition from long-term observations. In *Proc. European Conference on Computer Vision (ECCV)*, pages 851–865. Springer, 2002. 2

[33] N. Shroff, P. Turaga, and R. Chellappa. Moving vistas: Exploiting motion for describing scenes. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1911–1918. IEEE, 2010. 2, 6, 7

[34] C. Theriault, N. Thome, and M. Cord. Dynamic scene classification: Learning motion descriptors with slow features analysis. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2603–2610. IEEE, 2013. 7

[35] O. Tuzel, F. Porikli, and P. Meer. Pedestrian detection via classification on Riemannian manifolds. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 30(10):1713–1727, 2008. 3

[36] R. Wang, H. Guo, L. S. Davis, and Q. Dai. Covariance discriminative learning: A natural and efficient approach to image set classification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2496–2503. IEEE, 2012. 1, 6

[37] R. Wang, H. Guo, L. S. Davis, and Q. Dai. Covariance discriminative learning: A natural and efficient approach to image set classification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2496–2503. IEEE, 2012. 1, 2, 3, 5

[38] R. Wang, S. Shan, X. Chen, and W. Gao. Manifold-manifold distance with application to face recognition based on image set. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2008. 2

[39] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pages 487–495, 2014. 6