Beyond Gauss: Image-Set Matching on the Riemannian Manifold of PDFs

Mehrtash Harandi[1], Mathieu Salzmann[1,2], and Mahsa Baktashmotlagh[3]

[1]NICTA* and Australian National University, Canberra, Australia
[2]CVLab, EPFL, Switzerland
[3]Queensland University of Technology, Brisbane, Australia

Abstract

State-of-the-art image-set matching techniques typically implicitly model each image-set with a Gaussian distribution. Here, we propose to go beyond these representations and model image-sets as probability distribution functions (PDFs) using kernel density estimators. To compare and match image-sets, we exploit Csiszár f-divergences, which bear strong connections to the geodesic distance defined on the space of PDFs, i.e.*, the statistical manifold. Furthermore, we introduce valid positive definite kernels on the statistical manifold, which let us make use of more powerful classification schemes to match image-sets. Finally, we introduce a supervised dimensionality reduction technique that learns a latent space where f-divergences reflect the class labels of the data. Our experiments on diverse problems, such as video-based face recognition and dynamic texture classification, evidence the benefits of our approach over the state-of-the-art image-set matching methods.*

1. Introduction

This paper tackles the problem of image-set matching by comparing probability distribution functions (PDFs) using Csiszár f-divergences [9, 2]. Image-set matching, *i.e.* matching unordered sets of images, exploits the richer information contained in multiple images to perform recognition. With the growth of camera networks, video data, hyper-spectral imaging technologies, *etc.*, image-sets are becoming ubiquitous in our everyday lives.

State-of-the-art image-set matching methods [16, 43, 18, 24] typically model image-sets using geometrical structures, *e.g.*, Riemannian manifolds. The two most popular such structures are Grassmann manifolds [1] and the manifold of Symmetric Positive Definite (SPD) matrices [35]. From a different perspective, these representations can be related to modeling an image-set with a single multivariate Gaussian distribution. Indeed, in the case of the SPD manifold [43], an image-set is represented as the covariance matrix of the features extracted in each image of the set, which therefore essentially encodes a zero-mean Gaussian distribution of the image features. For Grassmann manifolds, where an image-set is represented by a subspace of its image features, several distances between subspaces were shown to be related to distances between multivariate Gaussian distributions [17]. While both manifolds have been shown to provide representations that are robust to varying imaging conditions, intuitively, modeling an image-set with a single Gaussian distribution seems restrictive.

In this paper, we therefore propose to make use of better probabilistic models to represent an image-set and study different ways to compare such models for the task of image-set matching. To this end, we model the PDF of an image-set using a non-parametric, data-driven kernel density estimator. Since PDFs form a Riemannian manifold, *i.e.*, the *statistical manifold*, the geodesic distance on the manifold comes as a natural choice to measure the similarity between two image-sets. Unfortunately, the geodesic distance is impractical to compute for general distributions. Therefore, we propose to exploit Csiszár f-divergences [9], which bear strong connections to the geodesic distance. In particular, we study two specific f-divergences and discuss how robust empirical estimates of these divergences can be obtained.

From a recognition perspective, f-divergences can be directly employed in a nearest neighbor classifier to match image-sets. However, for complex recognition tasks, a nearest neighbor classifier may have limited power. To address this issue, we therefore study the positive definiteness of kernels induced by these f-divergences.

In [25], Jaakkola and Haussler introduced a general form of positive definite kernels on statistical manifolds. How-

*NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the ARC through the ICT Centre of Excellence program.

ever, the kernels in [25] are characterized by the Fisher information matrix, and hence are only defined for statistical models that can be described by a finite set of parameters. By contrast, the kernels developed in this work relax this assumption and are positive definite for general distributions.

Finally, to account for the fact that the image features may not be the best representation to compare the data, we introduce a dimensionality reduction approach that exploits the f-divergences between image-sets. More specifically, we search for a low-dimensional space where the f-divergence between image-sets from different classes is maximized, while the f-divergence between image-sets of the same class is minimized. We then show that, thanks to a property of the f-divergences, dimensionality reduction can be cast as a minimization problem on a Grassmann manifold.

Our contributions can be summarized as: **1.** We introduce a new representation of image-sets as non-parametric PDFs. **2.** We propose to make use of f-divergences to perform image-set matching. **3.** We introduce a family of positive definite kernels on the space of PDFs. **4.** We derive a supervised dimensionality reduction technique that maximizes a notion of discriminative power between PDFs.

We evaluate the different algorithms derived from our analysis on several image-set matching tasks, including face recognition, dynamic texture categorization and scene classification. Our experiments evidence the benefits of our approach over state-of-the-art image-set matching methods. In particular, we outperform the state-of-the-art on the Youtube celebrity [27], DynTex++ [15], UCSD traffic [8] and Maryland scene recognition [40] datasets.

1.1. Related Work

As mentioned above, our work is motivated by methods that rely on geometrical structures, such as SPD and Grassmann manifolds, to represent image-sets, since they can be related to modeling image-sets with Gaussian distributions. We therefore focus our discussion on those methods.

In the context of SPD manifolds, Wang *et al*. proposed to model each image-set by a covariance matrix [43]. This allowed the authors to exploit the Riemannian geometry of SPD matrices to analyze image-sets. More specifically, [43] made use of a kernel function on the SPD manifold to perform kernel partial least squares regression or kernel discriminant analysis to recognize human faces in videos. Following this, [24] proposed to learn a combination of various kernels, including kernels on SPD manifolds, to boost the recognition accuracy.

The use of subspaces to match image-sets can be traced back to [44]. The main idea is to fit a subspace to the samples of an image-set and utilize the distance between multiple subspaces for classification. With the advance of methods that exploit the geometry of Grassmann manifolds, more sophisticated classifiers have been employed. For instance, Hamm and Lee [16] proposed to embed the Grassmannian into a Reproducing Kernel Hilbert Space (RKHS) and perform discriminant analysis in the resulting space. In [18], notions of sparse coding on the Grassmann manifold were utilized to perform image-set classification.

Since a covariance matrix inherently encodes a single Gaussian, and since several metrics between subspaces have been shown to be equivalent to distances between Gaussian distributions [17], both covariance-based and subspace-based representations can be related to modeling an image-set with a single multivariate Gaussian distribution. Note that Arandjelovic *et al*. [4] proposed to go beyond a single Gaussian by exploiting a Gaussian mixture model (GMM) in a low-dimensional space to represent an image-set. The similarity between two GMMs was then computed by a Monte-Carlo method. Here, in contrast, we make use of non-parametric density estimation together with robust empirical estimates of the distance between two PDFs. Furthermore, we define valid kernels on the statistical manifold and learn a low-dimensional representation that implicitly accounts for the f-divergence between the PDFs.

Of course, many other image-set matching techniques have been proposed in the past [32, 19, 26]. Notable examples include, but are not limited to, methods based on deep networks [21, 29], sparse coding and dictionary learning [28] and metric learning [42, 31]. While discussing these approaches in details goes beyond the scope of this paper, our experiments, which compare our results with the state-of-the-art in each dataset, demonstrate the benefits of our approach over these baselines.

2. Statistical Manifolds and f-Divergences

In this paper, we rely on probability density functions (PDFs) and on the distances between them to analyze image-sets. In this section, we therefore review some concepts related to the geometry of the space of PDFs.

Let $\mathcal{X}$ be a set. A PDF on $\mathcal{X}$ is a function $p : \mathcal{X} \to \mathbb{R}^+$ such that $\int_{\mathcal{X}} p(x)dx = 1$. Let $\mathcal{M}$ be a family of PDFs on the set $\mathcal{X}$. With certain assumptions (*e.g*., differentiability), $\mathcal{M}$ forms a Riemannian structure, *i.e*., a differentiable manifold equipped with a Riemannian metric. The Riemannian metric enables us to measure the length of curves[1].

The Fisher-Rao metric [37] is undoubtedly the most common Riemannian metric to analyze $\mathcal{M}$. Unfortunately, an analytic form of the geodesic distance induced by the Fisher-Rao metric can only be obtained for specific distributions, such as Gaussians, or the exponential family [37][2]. In other words, estimating geodesic distances between gen-

[1] On a Riemannian manifold, the geodesic distance between two points is the length of the shortest path on the manifold between them.

[2] Not to be confused with exponential distributions, or distributions generally expressed as sums of exponentials.

eral distributions, such as the ones we use here, is impractical. To address this issue, here, we propose to compare PDFs with two f-divergences, which, as discussed later, have strong connections with the geodesic distance.

Formally, a Csiszár f-divergence is a function of two probability distributions that measures their similarity, and is defined as

$$\delta_f(p\|q) = \int f\left(\frac{p}{q}\right) dq \ , \tag{1}$$

where f is a convex function on $(0, \infty)$ with $f(1) = 0$. Different choices of f yield different divergences. Below, and in the rest of this paper, we focus on two special cases, which induce the Hellinger distance and the Jeffrey divergence, respectively.

The Hellinger distance can be obtained by choosing $f(t) = (\sqrt{t} - 1)^2$ in Eq. 1, and is defined below.

Definition 1. *The Hellinger distance between two probability distributions p and q is defined as*

$$\delta_H^2(p\|q) = \int \left(\sqrt{p(\boldsymbol{x})} - \sqrt{q(\boldsymbol{x})}\right)^2 d\boldsymbol{x} \ . \tag{2}$$

If, instead, we set $f(t) = t\ln(t) - \ln(t)$ in Eq. 1, we obtain the Jeffrey divergence defined below.

Definition 2. *The Jeffrey or symmetric KL divergence between two probability distributions p and q is defined as*

$$\delta_J(p\|q) = \int \Big(p(\boldsymbol{x}) - q(\boldsymbol{x})\Big) \ln \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} d\boldsymbol{x} \ . \tag{3}$$

From a geometrical point of view, the Riemannian structure induced by the Hellinger distance is different from the one induced by the Jeffrey, or J-, divergence. However, these two divergences share the property that their respective Riemannian metrics can be obtained from the Fisher-Rao metric (see Thereom 5 in [3]). Furthermore, in [5], it was shown that the length of any given curve is the same under the Hellinger distance and under the Fisher-Rao metric up to scale. These two properties therefore relate these divergences to the geodesic distance on the statistical manifold, and thus make them an attractive alternative to compare PDFs. Another important property of these two divergences is given by the following theorem.

Theorem 2.1. *The Hellinger distance and the Jeffrey divergence between two distributions are invariant under differentiable and invertible transformations (diffeomorphisms).*

In other words, given two distributions $p_1(\boldsymbol{x})$ and $p_2(\boldsymbol{x})$ in a space $\mathcal{X}$, with $\boldsymbol{x} \in \mathcal{X}$, let $h : \mathcal{X} \to \mathcal{Y}$ be a differentiable and invertible function that maps $\boldsymbol{x}$ to $\boldsymbol{y}$. Under h, we have $p_i(\boldsymbol{x})d\boldsymbol{x} = q_i(\boldsymbol{y})d\boldsymbol{y}; i \in \{1, 2\}$ and $d(\boldsymbol{y}) = |\mathcal{J}(\boldsymbol{x})|d\boldsymbol{x}$, where $|\mathcal{J}(\boldsymbol{x})|$ denotes the determinant of the Jacobian matrix of h. The above invariance property states that

$$\delta_f(p_1(\boldsymbol{x}), p_2(\boldsymbol{x})) = \delta_f(q_1(\boldsymbol{y}), q_2(\boldsymbol{y})).$$

The proof of this theorem can be found in several recent studies (*e.g.*, Theorem 1 in [36]). It has also been known to mathematicians for decades [2]. Invariance to diffeomorphism seems an attractive property in the context of computer vision, and in particular for image-set matching, since images in a set can typically be subject to many variations, such as changes of illumination or of environment/capture conditions. Furthermore, we will exploit this property when deriving our dimensionality reduction method in Section 5. Note that the affine invariance of some metrics on the SPD manifold, which has made such metrics popular, is a lesser form of this property. In other words, the f-divergences are invariant to a broader set of transformations.

3. Image-Sets as PDFs

We now introduce our approach to modeling image-sets as PDFs. To this end, let $\{\boldsymbol{x}_i\}_{i=1}^n$ be a set of n images, where each $\boldsymbol{x}_i \in \mathbb{R}^D$ is a feature vector describing one image in the set. We propose to make use of Kernel Density Estimation (KDE) to obtain a fine-grained estimate $\hat{p}(\boldsymbol{x})$ of the distribution $p(\boldsymbol{x})$ of the features. In particular, we make use of Gaussian RBF kernels[3], which lets us write

$$\hat{p}(\boldsymbol{x}) = \frac{1}{n\sqrt{\det(2\pi\Sigma)}} \sum_{i=1}^{n} \exp\left(-\frac{1}{2}\Big(\boldsymbol{x} - \boldsymbol{x}_i\Big)^T \Sigma^{-1} \Big(\boldsymbol{x} - \boldsymbol{x}_i\Big)\right) . \tag{4}$$

Given two image-sets $\{\boldsymbol{x}_i^{(p)}\}_{i=1}^{n_p}$ and $\{\boldsymbol{x}_i^{(q)}\}_{i=1}^{n_q}$, Eq. 4 provides us with the means to estimate their respective PDFs $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$. We then aim to compare these image-sets by computing the statistical distance between $\hat{p}(\boldsymbol{x})$ and $\hat{q}(\boldsymbol{x})$. As discussed in Section 2, we propose to rely on f-divergences to compare $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$. Note, however, that the integrals corresponding to the Hellinger distance and to the J-divergence do not have an analytic solution for our KDE representation. Therefore, below, we derive solutions to compute a robust estimate of these two divergences.

3.1. Empirical f-Divergences

Let us first consider the case of the Hellinger distance. Given two sets of samples $\{\boldsymbol{x}_i^{(p)}\}_{i=1}^{n_p}$ and $\{\boldsymbol{x}_i^{(q)}\}_{i=1}^{n_q}$ drawn from $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$, respectively, directly estimating the integral of Eq. 2 is not straightforward. To make this easier, one can rewrite Eq. 2 as

$$\delta_H^2(p\|q) = \int \left(1 - \sqrt{\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}}\right)^2 p(\boldsymbol{x}) d\boldsymbol{x} \tag{5}$$

$$= E_p\left(1 - \sqrt{\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}}\right)^2 , \tag{6}$$

[3]Note that, in general, other kernels can be employed.

where $E_p(\cdot)$ denotes the expectation under p. Following the strong law of large numbers, as is commonly done in practice, such an expectation can then be estimated as

$$\hat{\delta}^2_H(p\|q) = \frac{1}{n_p}\sum_i^{n_p}\left(1 - \sqrt{\frac{\hat{q}(\boldsymbol{x}_i^{(p)}))}{\hat{p}(\boldsymbol{x}_i^{(p)})}}\right)^2, \tag{7}$$

with $\hat{p}(\cdot)$ and $\hat{q}(\cdot)$ obtained by KDE. Unfortunately, such an estimate would in general be different if one had chosen to make use of $q(\boldsymbol{x})$ instead of $p(\boldsymbol{x})$ to derive the expectation of Eq. 6. This implies that the resulting estimate of the Hellinger distance would be asymmetric, and thus poorly-suited to our goals.

Here, instead, we follow the approach of [7] to obtain a more robust estimate of the Hellinger distance. More specifically, we rewrite $\delta^2_H(p\|q)$ as

$$\begin{aligned}\delta^2_H(p\|q) &= \int\left(\sqrt{\frac{p(\boldsymbol{x})}{p(\boldsymbol{x})+q(\boldsymbol{x})}} - \sqrt{\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})+q(\boldsymbol{x})}}\right)^2 (p(\boldsymbol{x})+q(\boldsymbol{x}))d\boldsymbol{x} \\ &= E_p\left(\sqrt{T(\boldsymbol{x})} - \sqrt{1-T(\boldsymbol{x})}\right)^2 + E_q\left(\sqrt{T(\boldsymbol{x})} - \sqrt{1-T(\boldsymbol{x})}\right)^2,\end{aligned} \tag{8}$$

with

$$T(\boldsymbol{x}) = \frac{p(\boldsymbol{x})}{p(\boldsymbol{x})+q(\boldsymbol{x})}. \tag{9}$$

Given our two sets of samples $\{\boldsymbol{x}_i^{(p)}\}_{i=1}^{n_p}$ and $\{\boldsymbol{x}_i^{(q)}\}_{i=1}^{n_q}$, this allows us to estimate the Hellinger distance as

$$\begin{aligned}\hat{\delta}^2_H(p\|q) = &\frac{1}{n_p}\sum_i^{n_p}\left(\sqrt{T(\boldsymbol{x}_i^{(p)})} - \sqrt{1-T(\boldsymbol{x}_i^{(p)})}\right)^2 \\ &+ \frac{1}{n_q}\sum_i^{n_q}\left(\sqrt{T(\boldsymbol{x}_i^{(q)})} - \sqrt{1-T(\boldsymbol{x}_i^{(q)})}\right)^2.\end{aligned} \tag{10}$$

The benefits of this approach are twofold. First, the resulting estimate is symmetric. Second, and maybe more importantly, the denominator of $T(\cdot)$ alleviates the potential instabilities that low probabilities of samples under either $\hat{q}$ or $\hat{p}$ would have resulted in by making use of the formulation in Eq. 7, or of its counterpart in terms of q. Note that such low probabilities are quite common when relying on KDE with high-dimensional data.

In the case of the Jeffrey divergence, we make use of the same idea as for the Hellinger distance. We therefore express the J-divergence in terms of $T(\cdot)$, which, after some derivations, yields

$$\begin{aligned}\hat{\delta}_J(p\|q) = &\frac{1}{n_p}\sum_i^{n_p}(2T(\boldsymbol{x}_i^{(p)})-1)\ln\frac{T(\boldsymbol{x}_i^{(p)})}{1-T(\boldsymbol{x}_i^{(p)})} \\ &+ \frac{1}{n_q}\sum_i^{n_q}(2T(\boldsymbol{x}_i^{(q)})-1)\ln\frac{T(\boldsymbol{x}_i^{(q)})}{1-T(\boldsymbol{x}_i^{(q)})}.\end{aligned} \tag{11}$$

Our two empirical estimates give us practical ways to evaluate the distance between two image-sets represented by their PDFs. Given a training set of m image-sets and a query image-set, matching can then simply be achieved by computing the distance of the query to all training image-sets, and choosing the nearest one as matching set.

4. Kernels on the Statistical Manifold

In the previous section, we have introduced an approach to comparing the distributions of two image-sets using empirical estimates of the Hellinger distance or of the J-divergence. Such an approach, however, only allows us to make use of a nearest neighbor classifier. Kernel methods (*e.g.*, Kernel Fisher discriminant analysis), however, provide much more powerful tools to perform classification, and thus image-set matching. Here, we therefore turn to the question of whether the divergences defined in Section 2 can generate valid positive definite (*pd*) kernels. To answer this question, let us first define the notion of pd kernels.

Definition 3 (Real-valued Positive Definite Kernels)**.** *Let $\mathcal{X}$ be a nonempty set. A symmetric function $k : \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ is a positive definite (***pd***) kernel on $\mathcal{X}$ if and only if $\sum_{i,j=1}^n c_i c_j k(x_i,x_j) \geq 0$ for any $n\in\mathbb{N}$, $x_i\in\mathcal{X}$ and $c_i\in\mathbb{R}$.*

For the J-divergence, the kernel

$$k_J(p,q) = \exp\left(-\sigma\delta_J(p\|q)\right) \tag{12}$$

was introduced in [33], although without a formal proof of positive definiteness. To the best of our knowledge, a counter-example that disproves the positive definiteness of $k_J(\cdot,\cdot)$ has never been exhibited in the literature. Therefore, in our experiments, we assumed that $k_J(\cdot,\cdot)$ is *pd*. Note that, in contrast to the Hellinger distance, square root(Jeffrey divergence) is not a metric, as can be shown by a counter example. This motivated our notation δ_J instead of δ^2_J. However, since δ_J is the counterpart of δ^2_H for a different function f in the definition of the f-divergence, the kernel in Eq.12 can still be thought of as a Gaussian kernel.

In the case of the Hellinger distance, a conditionally positive definite (*cpd*) kernel was studied in [22]. Here, in contrast, we derive valid *pd* kernels. More precisely, we do not only introduce a single *pd* kernel, but rather provide a recipe to generate diverse *pd* kernels on the statistical manifold. Our derivations rely on the definition of negative definite kernels given below.

Definition 4 (Real-valued Negative Definite Kernels)**.** *Let $\mathcal{X}$ be a nonempty set. A symmetric function $\psi : \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ is a negative definite (***nd***) kernel on $\mathcal{X}$ 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 $\sum_{i=1}^n c_i = 0$.*

Note that, in contrast to positive definite kernels, an additional constraint of the form $\sum c_i = 0$ is required in the negative definite case. Given this definition, we now prove that the Hellinger distance is *nd*.

Theorem 4.1 (Negative Definiteness of the Hellinger distance)**.** *Let $\mathcal{M}$ denote the statistical manifold. The Hellinger distance $\delta_H^2 : \mathcal{M} \times \mathcal{M} \to \mathbb{R}^+$ is negative definite.*

Proof.

$$\sum_{i,j=1}^{N} c_i c_j \delta_H^2(p_i \| p_j) = \sum_{i,j=1}^{N} c_i c_j \int_x \left(\sqrt{p_i(x)} - \sqrt{p_j(x)}\right)^2 dx$$

$$= \sum_{i=1}^{N} c_i \sum_{j=1}^{N} c_j \int_x p_j(x)dx + \sum_{j=1}^{N} c_j \sum_{i=1}^{N} c_i \int_x p_i(x)dx$$

$$- 2 \sum_{i,j=1}^{N} c_i c_j \int_x \sqrt{p_i(x)p_j(x)}dx$$

$$= -2 \int_x \sum_{i}^{N} c_i \sqrt{p_i(x)} \sum_{j}^{N} c_j \sqrt{p_j(x)}dx$$

$$= -2 \int_x \| \sum_{i}^{N} c_i \sqrt{p_i(x)} \|^2 dx \leq 0 \,,$$

where the terms in the second line have disappeared due to the constraints $\sum_i c_i = 0$, resp. $\sum_j c_j = 0$, and to the fact that the integrals have value 1 for any i, resp. j. □

We then make use of the following theorem, which originated from the work of I. J. Schoenberg (1903-1990).

Theorem 4.2 (Theorem 2.3 in Chapter 3 of [6])**.** *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 $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ is negative definite.*

Theorem 4.2 provides a general recipe to create *pd* kernels. In particular, here, we focus on the Gaussian and the Laplace kernels, which have proven powerful in Euclidean space. The Gaussian kernel can be obtained by choosing $\mu(t) = \delta(t-1)$ in Theorem 4.2, where δ denotes the Dirac function. On the statistical manifold, this kernel can then be written as

$$k_H(p,q) = \exp(-\sigma \delta_H^2(p,q)), \quad \sigma > 0 \,. \tag{13}$$

To derive the Laplace kernel on the statistical manifold, we must further rely on the following theorem.

Theorem 4.3 (Corollary 2.10 in Chapter 3 of [6])**.** *If $\psi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is negative definite and satisfies $\psi(\boldsymbol{x}, \boldsymbol{x}) \geqq 0$ then so is ψ^α for $0 < \alpha < 1$.*

As a consequence, by choosing $\psi = \delta_H^2$ and $\alpha = 1/2$ in Theorem 4.3, we have that $\delta_H(\cdot,\cdot)$ is *nd*. Then, applying Theorem 4.2 with $\delta_H(\cdot,\cdot)$ and $\mu(t) = \delta(t-1)$ lets us derive the Laplace kernel on the statistical manifold

$$k_L(p,q) = \exp(-\sigma \delta_H(p,q)), \quad \sigma > 0 \,. \tag{14}$$

Remark 1. *The Hellinger distance can be thought of as the chordal distance between points on an infinite-dimensional unit hyper-sphere. More specifically, the square root function is a diffeomorphism between the statistical manifold and the unit hyper-sphere. In [41], this was exploited to estimate the distance between discretized PDFs as the geodesic distance on the corresponding (finite-dimensional) hyper-sphere. Such a distance, however, cannot induce a valid positive definite Gaussian kernel, since the Gaussian kernel produced by the geodesic distance on a Riemannian manifold is not positive definite unless the manifold is flat [14]. In contrast, as shown above, our divergences yield valid positive definite kernels, which allows us to exploit more sophisticated classification methods.*

Remark 2. *Note that the discussion above proves the positive definiteness of kernels defined with the exact Hellinger distance, and does not necessarily extend to its empirical estimate. However, since the strong law of large numbers guarantees convergence of our empirical estimate to the true distance, given sufficiently many samples, the resulting empirical kernels will also be* pd.

In our experiments, we used the kernels derived above to perform kernel discriminant analysis. Image-set matching was then achieved by using the Euclidean distance in the resulting low-dimensional latent space.

5. f-Divergences for Dimensionality Reduction

The methods described in Sections 3 and 4 directly compare the distributions of the original features of each image-set. As mentioned earlier, with high-dimensional features that are common in computer vision, KDE may produce very sparse PDFs (*i.e.*, PDFs that are strongly peaked around the samples and zero everywhere else), which may be less reliable to compare. To address this issue, here, we propose to learn a mapping of the features to a low-dimensional space, such that the f-divergences in the resulting space reflect some interesting properties of the data.

As shown below, we formulate dimensionality reduction as an optimization problem on a Grassmann manifold. The use of Grassmann and Stiefel manifolds for dimensionality reduction is an emerging topic in machine learning. Two notable examples are robust PCA using the Grassmannian [20] and linear dimensionality reduction using Stiefel manifolds [10].

Focusing on the supervised scenario, we search for a latent space where two image-sets are close to each other

(according to the f-divergence) if they belong to the same class and far apart if they don't. That is, given a set of image-sets $\mathcal{X} = \{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_m\}$, where each image-set $\boldsymbol{X}_i = \{\boldsymbol{x}_l^{(i)}\}_{l=1}^{n_i}$, $\boldsymbol{x}_l^{(i)} \in \mathbb{R}^D$, our goal is to find a transformation $\boldsymbol{W} \in \mathbb{R}^{D\times d}$ such that the f-divergences between the mapped image-sets $\left\{\{\boldsymbol{W}^T\boldsymbol{x}_l^{(i)}\}_{l=1}^{n_i}\right\}_{i=1}^{m}$ encode some interesting structure of the data. Here, we represent this structure via an affinity function $a(\boldsymbol{X}_i, \boldsymbol{X}_j)$ that encodes pairwise relationships between the image-sets. This affinity function will be described in Section 5.1.

Since we aim for the f-divergences to reflect this affinity measure, we can write a cost function of the form

$$\mathcal{L}(\boldsymbol{W}) = \sum_{i,j} a(\boldsymbol{X}_i, \boldsymbol{X}_j) \cdot \delta\Big(\boldsymbol{W}^T\boldsymbol{X}_i, \boldsymbol{W}^T\boldsymbol{X}_j\Big) \ , \quad (15)$$

where $\delta : \mathcal{M} \times \mathcal{M} \to \mathbb{R}^+$ is either $\delta_H^2(\cdot,\cdot)$ or $\delta_J(\cdot,\cdot)$, and where we sum over all pairs of image-sets. To avoid possible degeneracies when minimizing this cost function w.r.t. $\boldsymbol{W}$, and following common practice in dimensionality reduction, we enforce the solution to be orthogonal, *i.e.*, $\boldsymbol{W}^T\boldsymbol{W} = \mathbf{I}_d$. This allows us to write dimensionality reduction as the optimization problem

$$\begin{aligned} \boldsymbol{W}^* = \quad & \underset{\boldsymbol{W}}{\operatorname{argmin}} \ \mathcal{L}(\boldsymbol{W}) \\ & \text{s.t. } \boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{I}_d \ . \end{aligned} \quad (16)$$

Below, we show that, for both divergences of interest, *i.e.*, the Hellinger distance and the J-divergence, (16) is a minimization problem on a Grassmann manifold.

The Grassmann manifold $\mathcal{G}(d, D)$ is the space of d-dimensional subspaces in $\mathbb{R}^D$ and corresponds to as a quotient space of the Stiefel manifold (*i.e.*, the space of d-dimensional frames in $\mathbb{R}^D$, or in other words orthogonal $D \times d$ matrices) [12]. More specifically, the points on the Stiefel manifold that form a basis of the same subspace are identified with a single point on the Grassmann manifold. As such, a minimization problem with orthogonality constraint $\boldsymbol{W}^T\boldsymbol{W} = \mathbf{I}_d$ is a problem on the Grassmannian *iff* the cost of the problem is invariant to the choice of basis of the subspace spanned by $\boldsymbol{W}$. Mathematically, $\min_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{W})$ with $\boldsymbol{W}^T\boldsymbol{W} = \mathbf{I}_d$ is a problem on the Grassmannian *iff* $\mathcal{L}(\boldsymbol{W}\boldsymbol{R}) = \mathcal{L}(\boldsymbol{W})$, $\forall \boldsymbol{R} \in \mathcal{O}(d)$, where $\mathcal{O}(d)$ denotes the group of $d \times d$ orthogonal matrices. Since transformations in $\mathcal{O}(d)$ are bijections, the invariance property of Theorem 2.1 directly shows that the cost function of (16) is invariant to the choice of basis. In other words, (16) can be solved as an unconstrained minimization problem on $\mathcal{G}(d, D)$.

In practice, to solve (16) on $\mathcal{G}(d, D)$, we make use of Newton-type methods (*e.g.*, the conjugate gradient method). These methods inherently require the gradient of $\mathcal{L}(\boldsymbol{W})$. On $\mathcal{G}(d, D)$, the gradient can be expressed as

$$\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W}) = (\mathbf{I}_D - \boldsymbol{W}\boldsymbol{W}^T)\,\operatorname{grad}\mathcal{L}(\boldsymbol{W}), \quad (17)$$

where $\operatorname{grad}\mathcal{L}(\boldsymbol{W})$ is the $D \times d$ matrix of partial derivatives of $\mathcal{L}(\boldsymbol{W})$ with respect to the elements of $\boldsymbol{W}$, *i.e.*,

$$[\operatorname{grad}\mathcal{L}(\boldsymbol{W})]_{i,j} = \frac{\partial\mathcal{L}(\boldsymbol{W})}{\partial\boldsymbol{W}_{i,j}}.$$

The detailed derivations of $\operatorname{grad}\mathcal{L}(\boldsymbol{W})$ for our two f-divergences are provided in supplementary material.

5.1. Affinity Measure

As mentioned above, we propose to exploit supervised data to define the affinity measure used in the cost function of Eq. 15. Note that unsupervised approaches are also possible, for instance to find a mapping that preserves the closeness of pairs of image-sets.

In our supervised scenario, let y_i denote the class label of image-set $\boldsymbol{X}_i$, with $1 \le y_i \le C$. We define the affinity between two sets $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ with labels y_i and y_j, respectively, as

$$a(\boldsymbol{X}_i, \boldsymbol{X}_j) = g_w(\boldsymbol{X}_i, \boldsymbol{X}_j) - g_b(\boldsymbol{X}_i, \boldsymbol{X}_j) \ , \quad (18)$$

where g_w and g_b encode a notion of within-class similarity and between-class similarity, respectively. These similarities can be expressed as

$$g_w(\boldsymbol{X}_i, \boldsymbol{X}_j) = \begin{cases} 1, & \text{if } \boldsymbol{X}_i \in N_w(\boldsymbol{X}_j) \text{ or } \boldsymbol{X}_j \in N_w(\boldsymbol{X}_i) \\ 0, & \text{otherwise} \end{cases}$$

$$g_b(\boldsymbol{X}_i, \boldsymbol{X}_j) = \begin{cases} 1, & \text{if } \boldsymbol{X}_i \in N_b(\boldsymbol{X}_j) \text{ or } \boldsymbol{X}_j \in N_b(\boldsymbol{X}_i) \\ 0, & \text{otherwise} \end{cases}$$

where $N_w(\boldsymbol{X}_i)$ is the set of ν_w nearest neighbors of $\boldsymbol{X}_i$ (according to the f-divergence) that share the same label as y_i, and $N_b(\boldsymbol{X}_i)$ contains the ν_b nearest neighbors of $\boldsymbol{X}_i$ having different labels. In our experiments, we defined ν_w as the minimum number of points in a class and found $\nu_b \le \nu_w$ by cross-validation.

Before presenting our complete set of experiments, we would like to provide some insights regarding our dimensionality reduction algorithm. The examples shown below are all taken from the face recognition experiment on the YouTube Celebrity (YTC) dataset [27] (the first experiment in Section 6). First, in Fig. 1, we illustrate the convergence of (16) optimized using a conjugate gradient method on the Grassmannian. In practice, we found that the algorithm typically converges in less than 25 iterations. For YTC, each conjugate gradient iteration took roughly 40 seconds on a quad core desktop machine. Second, in Fig. 2, we provide the matrices of pairwise f-divergences, before and after dimensionality reduction, for samples taken from eight representative classes of the YTC dataset. Bright and dark regions represent high and low similarities, respectively. The

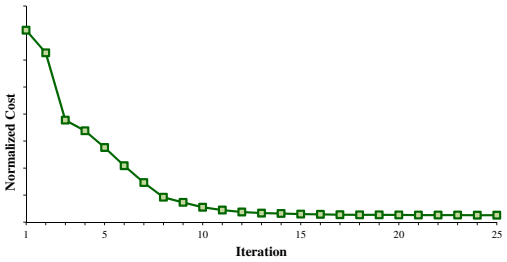

Figure 1: Convergence of (16) using a conjugate gradient method on the Grassmannian for the YTC dataset.

ideal affinity matrix should contain eight 3×3 blocks on its diagonal. Fig. 2 clearly evidences that our dimensionality reduction procedure yields a matrix that better matches the ideal affinity matrix. A further benefit of dimensionality reduction is the gain in speed to compute divergences. For example, in the case of YTC, computing 10,000 distances in the high-dimensional space took 100 seconds vs. 25 seconds after dimensionality reduction.

6. Experimental Evaluation

We now evaluate the algorithms introduced in the previous sections on diverse standard image-set matching problems. In particular, for our kernel-based approach, we make use of the kernel Fisher Discriminant Analysis (kFDA) algorithm. kFDA is a kernel-based approach to learning a discriminative latent space. Classification in the resulting latent space is then performed with a Nearest Neighbor (NN) classifier based on the Euclidean distance. In all our experiments, the dimensionality of the latent space, for both kFDA and our dimensionality reduction scheme, was determined by cross-validation. The kernel bandwidth, *i.e.*, σ in Eq.12, Eq.13 and Eq.14 was chosen by cross-validation from the set $\{0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1\}$. In the remainder of this section, we refer to our different algorithms as

NN-H: NN classifier based on the Hellinger distance.

NN-J: NN classifier based on the J-divergence.

kFDA-HG: kFDA with the Hellinger distance (Eq. 13).

kFDA-HL: kFDA with the Hellinger distance (Eq. 14).

kFDA-J: kFDA with the J-divergence (Eq. 12).

NN-H-DR: NN classifier based on the Hellinger distance after our dimensionality reduction scheme.

NN-J-DR: NN classifier based on the J-divergence after our dimensionality reduction scheme.

Since our approach was motivated by techniques that exploit geometrical structures, such as SPD or Grassmann manifolds, which have proven effective for image-set matching, we compare our results against two such techniques. In particular, we make use of Grassmann Discriminant Analysis (**GDA**) [16] and of Covariance Discriminative Learning (**CDL**) [43] as baseline algorithms, both of which, as us, employ kFDA to match image-sets. For CDL, the kernel function $k_{\mathcal{S}} : \mathcal{S}^n_{++} \times \mathcal{S}^n_{++} \rightarrow \mathbb{R}$ is given by $k_{\mathcal{S}}(\boldsymbol{A}, \boldsymbol{B}) = \mathrm{Tr}(\log(\boldsymbol{A})^T \log(\boldsymbol{B}))$, where log is the principal matrix logarithm. For GDA, the kernel function $k_{\mathcal{G}} : \mathcal{G}(p,n) \times \mathcal{G}(p,n) \rightarrow \mathbb{R}$ is the projection kernel defined as $k_{\mathcal{G}}(\boldsymbol{A}, \boldsymbol{B}) = \|\boldsymbol{A}^T \boldsymbol{B}\|^2_F$. In all our experiments, we used the same image features for our approach and for GDA and CDL. Finally, for each dataset, we also report the best result found in the literature, and refer to this result as **State-of-the-art**.

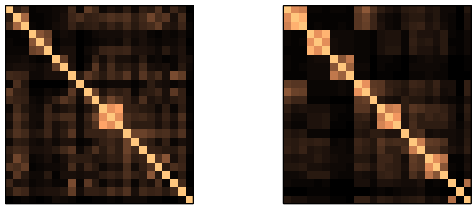

Figure 2: Matrix of pairwise f-divergences before and after dimensionality reduction for 8 classes of YTC. Note that the matrix obtained after dimensionality reduction better matches the ideal matrix, which should contain eight 3×3 blocks on its diagonal.

6.1. Video-Based Face Recognition

For the task of image-set-based face recognition, we used the YTC [27] and COX [23] datasets. The YTC dataset contains 1910 video clips of 47 subjects. The COX dataset includes 1,000 subjects, each captured by three cameras (*i.e.*, 3,000 videos in total). For the YTC dataset, we described each region with a histogram of Local Binary Patterns (LBP) [34]. For the COX dataset, following [24], we used histograms of equalized intensity values as features, which were found to be superior to LBP features on COX.

For the YTC dataset, following the standard practice [30], 3 videos from each person were randomly chosen as training/gallery data, and the query set contained 6 randomly chosen videos from each subject. The process of random selection was repeated 5 times. For the COX dataset, we followed the test protocol of [24]: 100 subjects were randomly chosen to form the gallery/probe sets for 6 different experiments. For each experiment, the camera number determines the gallery and probe sets. For example, COX12 refers to the test scenario where videos captured by Cam1 and Cam2 are used as gallery and probe set, respectively. The random selection of training and gallery/probe sets was repeated 10 times.

Table 1 shows the average accuracies of all methods on the YTC and COX datasets. For these datasets, the state-of-

Table 1: Average recognition rates on the YTC, COX, DynTex++, UCSD traffic and Maryland scene (ML) datasets.

Method	YTC	COX12	COX13	COX23	COX21	COX31	COX32	DynTex++	Traffic	ML-LOO	ML
GDA [16]	66.2 ± 9.7	68.8	77.7	71.6	66.0	76.1	74.8	89.9 ± 0.6	92.5 ± 2.6	81.5	70.3 ± 5.2
CDL [43]	70.9 ± 3.2	78.4	85.3	79.7	75.6	85.8	81.9	89.0 ± 0.9	91.7 ± 1.9	86.5	76.7 ± 7.8
State-of-the-art	78.2	**95.1**	96.3	94.2	92.3	95.4	**94.5**	93.8	95.6	77.7	NA
NN-H	77.3 ± 4.5	61.7 ± 4.2	69.2 ± 4.0	63.5 ± 2.3	66.6 ± 5.0	64.2 ± 4.2	64.0 ± 3.5	91.6 ± 0.7	91.3 ± 4.2	76.9	71.2 ± 3.1
NN-J	76.7 ± 5.1	64.7 ± 4.1	69.5 ± 3.3	63.0 ± 2.4	65.5 ± 5.1	70.0 ± 3.9	63.3 ± 3.6	91.4 ± 0.7	91.0 ± 4.5	77.7	71.4 ± 3.0
kFDA-HG	78.9 ± 3.4	90.8 ± 3.0	96.0 ± 1.7	92.9 ± 1.9	$\mathbf{92.5 \pm 2.4}$	95.8 ± 1.7	93.4 ± 1.7	94.7 ± 0.4	96.1 ± 1.5	85.4	78.1 ± 4.4
kFDA-HL	78.6 ± 4.7	92.4 ± 2.1	$\mathbf{96.8 \pm 0.7}$	$\mathbf{94.7 \pm 1.1}$	92.2 ± 1.1	$\mathbf{96.6 \pm 0.8}$	93.7 ± 1.3	94.9 ± 0.7	96.5 ± 1.5	**87.7**	79.0 ± 3.1
kFDA-J	$\mathbf{79.4 \pm 3.8}$	91.5 ± 3.0	95.9 ± 1.7	93.0 ± 2.0	92.5 ± 2.5	95.6 ± 1.5	93.5 ± 1.6	$\mathbf{95.2 \pm 0.6}$	$\mathbf{97.3 \pm 1.4}$	86.9	77.8 ± 5.3
NN-H-DR	78.3 ± 3.7	71.1 ± 4.0	83.6 ± 3.5	77.1 ± 4.1	76.6 ± 3.6	76.4 ± 4.5	77.1 ± 3.5	92.3 ± 0.5	94.9 ± 2.9	80.8	79.7 ± 4.5
NN-J-DR	79.3 ± 3.6	72.3 ± 2.9	82.6 ± 3.3	75.6 ± 3.1	73.2 ± 3.1	81.8 ± 2.7	70.4 ± 3.8	91.9 ± 0.5	95.6 ± 1.5	82.3	$\mathbf{80.2 \pm 3.7}$

the-art baselines correspond to the metric learning approach of [30] and the hybrid solution of [24], respectively. As far as geometrical methods are concerned, the results evidence that making use of the statistical manifold yields superior results compared to the Grassmann and SPD manifolds. This is even true for the direct NN classifiers based on our divergences, which are further improved by dimensionality reduction. This, we believe, demonstrates the benefits of relying on more accurate PDF representations of each image-set (*i.e.*, KDE in our case versus single Gaussians for CDL and GDA). Furthermore, on both datasets, our algorithms either match or outperform the state-of-the-art. The exceptions are COX12 and COX32, which could be attributed to the more sophisticated classification scheme used in [24]. Note that, as acknowledged in [24], the hybrid method does not scale up well to large datasets. By contrast, and as evidenced by our results on the full COX dataset in supplementary material, our approach can easily handle large amounts of data.

6.2. Dynamic Texture Recognition

As a second task, we considered the problem of dynamic texture recognition using the DynTex++ dataset [15]. DynTex++ [15] is comprised of 36 classes, each of which contains 100 sequences.We split the dataset into training and test sets by randomly assigning half of the videos of each class to the training set and using the rest as query data. We used the LBP-TOP [45] approach to represent each video sequence. Table 1 shows the average accuracies for 10 random splits. To the best of our knowledge, [38] reported the highest accuracy on this dataset. Our kFDA-J approach yields a 1.4% improvement over this state-of-the-art. As before, we can observe a gap between our approach and GDA or CDL.

6.3. Scene Classification

For scene classification, we employed the UCSD traffic dataset [8] and the Maryland scene recognition dataset [40].

For UCSD, we used HoG features [11] to describe each frame. Our experiments were performed using the splits provided with the dataset [8]. The state-of-the-art results were reported in [39]. Once again, we see that our kernel-based and dimensionality reduction algorithms comfortably outperform GDA and CDL, as well as the previous state-of-the-art.

Figure 3: Representative examples of three classes of the Maryland scene dataset [40]. From left to right: iceberg collapsing, tornado, and volcano eruption.

As a last experiment, we used the Maryland dataset, which contains 13 different classes of dynamic scenes. This dataset is more challenging, and we observed that hand-crafted features, such as LBP or HoG, do not provide sufficiently discriminative representations. Therefore, we used the last layer of the CNN trained in [46] as frame descriptors. We then reduced the dimensionality of the CNN output to 400 using PCA. We first employed the standard Leave-One-Out (LOO) test protocol. Furthermore, we also evaluated the methods on 10 different training/query partitions obtained by randomly choosing 70% of the dataset for training and the remaining 30% for testing. The average classification accuracies for both protocols are reported in Table 1. Note that no state-of-the-art results have been reported in the literature on our second test protocol. Our approach outperforms the state-of-the-art result of Feichtenhofer [13] by more than 7%. While this may be attributed in part to the CNN features, note that our approach still outperforms GDA and CDL based on the same features.

7. Conclusions and Future Work

We have introduced a novel framework to model and compare image-sets. Specifically, we have made use of KDE to represent an image-set with its PDF, and have proposed practical solutions to employ f-divergences for image-set matching, including empirical estimates of f-divergences, valid *pd* kernels on the statistical manifold and a supervised dimensionality reduction algorithm inherently accounting for f-divergences in the resulting latent space. In the future, we plan to extend our learning scheme to the unsupervised and semi-supervised scenarios. Furthermore, we intend to study the use and effectiveness of other divergences to tackle the problem of image-set matching.

References

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

[2] S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1966.

[3] S.-I. Amari and A. Cichocki. Information geometry of divergence functions. *Bulletin of the Polish Academy of Sciences: Technical Sciences*, 58(1), 2010.

[4] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell. Face recognition with image sets using manifold density divergence. In *CVPR*, 2005.

[5] M. Baktashmotlagh, M. Harandi, B. Lovell, and M. Salzmann. Domain adaptation on the statistical manifold. In *CVPR*, 2014.

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

[7] K. M. Carter. *Dimensionality reduction on statistical manifolds*. PhD thesis, University of Michigan, 2009.

[8] A. B. Chan and N. Vasconcelos. Probabilistic kernels for the classification of auto-regressive visual processes. In *CVPR*, 2005.

[9] I. Csiszár. Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizitat von markoffschen ketten. *Magyar. Tud. Akad. Mat. Kutato Int. Kozl*, 8:85–108, 1963.

[10] J. P. Cunningham and Z. Ghahramani. Linear dimensionality reduction: Survey, insights, and generalizations. *JMLR*, pages –, 2015.

[11] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[12] A. Edelman, T. A. Arias, and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.*, 1999.

[13] C. Feichtenhofer, A. Pinz, and R. Wildes. Bags of spacetime energies for dynamic scene recognition. In *CVPR*, 2014.

[14] A. Feragen, F. Lauze, and S. Hauberg. Geodesic exponential kernels: When curvature and linearity conflict. In *CVPR*, pages 3032–3042, June 2015.

[15] B. Ghanem and N. Ahuja. Maximum margin distance learning for dynamic texture recognition. In *ECCV*, 2010.

[16] J. Hamm and D. D. Lee. Grassmann discriminant analysis: a unifying view on subspace-based learning. In *ICML*, 2008.

[17] J. Hamm and D. D. Lee. Extended Grassmann kernels for subspace-based learning. In *NIPS*, 2009.

[18] M. Harandi, R. Hartley, C. Shen, B. Lovell, and C. Sanderson. Extrinsic methods for coding and dictionary learning on Grassmann manifolds. *IJCV*, 114(2-3):113–136, 2015.

[19] M. Harandi, M. Salzmann, S. Jayasumana, R. Hartley, and H. Li. Expanding the family of grassmannian kernels: An embedding perspective. In *ECCV*, volume 8695, pages 408–423. 2014.

[20] S. Hauberg, A. Feragen, and M. J. Black. Grassmann averages for scalable robust pca. In *CVPR*, pages 3810–3817, 2014.

[21] M. Hayat, M. Bennamoun, and S. An. Deep reconstruction models for image set classification. *TPAMI*, 37(4), 2015.

[22] M. Hein and O. Bousquet. Hilbertian metrics and positive definite kernels on probability measures. In *AISTATS*, 2005.

[23] Z. Huang, S. Shan, H. Zhang, S. Lao, A. Kuerban, and X. Chen. Benchmarking still-to-video face recognition via partial and local linear discriminant analysis on cox-s2v dataset. In *ACCV*, 2013.

[24] Z. Huang, R. Wang, S. Shan, and X. Chen. Face recognition on large-scale video in the wild with hybrid euclidean and Riemannian metric learning. *Pattern Recognition (PR)*, 2015.

[25] T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *NIPS*, pages 487–493, 1999.

[26] S. Jayasumana, R. Hartley, M. Salzmann, H. Li, and M. Harandi. Kernel methods on riemannian manifolds with gaussian rbf kernels. *TPAMI*, 2015.

[27] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley. Face tracking and recognition with visual constraints in real-world videos. In *CVPR*, 2008.

[28] J. Lu, G. Wang, W. Deng, and P. Moulin. Simultaneous feature and dictionary learning for image set based face recognition. In *ECCV*. 2014.

[29] J. Lu, G. Wang, W. Deng, P. Moulin, and J. Zhou. Multi-manifold deep metric learning for image set classification. In *CVPR*, 2015.

[30] J. Lu, G. Wang, and P. Moulin. Image set classification using holistic multiple order statistics features and localized multi-kernel metric learning. In *ICCV*, 2013.

[31] A. Mian, Y. Hu, R. Hartley, and R. Owens. Image set based face recognition using self-regularized non-negative coding and adaptive distance metric learning. *TIP*, 22(12), 2013.

[32] H. Mobahi, C. Liu, and W. Freeman. A compositional model for low-dimensional image set representation. In *CVPR*, pages 1322–1329, 2014.

[33] P. J. Moreno, P. P. Ho, and N. Vasconcelos. A Kullback-Leibler divergence based kernel for svm classification in multimedia applications. In *NIPS*. 2004.

[34] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *TPAMI*, 24, 2002.

[35] X. Pennec, P. Fillard, and N. Ayache. A Riemannian framework for tensor computing. *IJCV*, 66(1), 2006.

[36] Y. Qiao and N. Minematsu. A study on invariance of-divergence and its application to speech recognition. *IEEE Trans. on Signal Processing*, 58(7), 2010.

[37] C. Radhakrishna Rao. Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37(3), 1945.

[38] A. Ramirez Rivera and O. Chae. Spatiotemporal directional number transitional graph for dynamic texture recognition. *TPAMI*, 2015.

[39] A. Ravichandran, P. Favaro, and R. Vidal. A unified approach to segmentation and categorization of dynamic textures. In *ACCV*. 2011.

[40] N. Shroff, P. Turaga, and R. Chellappa. Moving vistas: Exploiting motion for describing scenes. In *CVPR*, 2010.

[41] A. Srivastava, I. Jermyn, and S. Joshi. Riemannian analysis of probability density functions with applications in vision. In *CVPR*, 2007.

[42] R. Vemulapalli, J. K. Pillai, and R. Chellappa. Kernel learning for extrinsic classification of manifold features. In *CVPR*, 2013.

[43] R. Wang, H. Guo, L. S. Davis, and Q. Dai. Covariance discriminative learning: A natural and efficient approach to image set classification. In *CVPR*, 2012.

[44] O. Yamaguchi, K. Fukui, and K. Maeda. Face recognition using temporal image sequence. In *FG*, 1998.

[45] G. Zhao and M. Pietikäinen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *TPAMI*, 29(6), 2007.

[46] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *NIPS*, 2014.