Learning Deep Match Kernels for Image-Set Classification

Haoliang Sun[1,2], Xiantong Zhen[2], Yuanjie Zheng[3], Gongping Yang[1], Yilong Yin[1,4*], and Shuo Li[2]

[1]Shandong University, Jinan, China
[2]The University of Western Ontario, London, ON, Canada
[3]School of Information Science and Engineering, Key Lab of Intelligent Computing & Information Security in Universities of Shandong, Key Lab of Intelligent Information Processing, Institute of Life Sciences at Shandong Normal University, Jinan, China
[4]Shandong University of Finance and Economics, Jinan, China
{haolsun.cn,zhenxt,zheng.vision,slishuo}@gmail.com, {ylyin,gpyang}@sdu.edu.cn

Abstract

Image-set classification has recently generated great popularity due to its widespread applications in computer vision. The great challenges arise from effectively and efficiently measuring the similarity between image sets with high inter-class ambiguity and huge intra-class variability. In this paper, we propose deep match kernels (DMK) to directly measure the similarity between image sets in the match kernel framework. Specifically, we build deep local match kernels between images upon arc-cosine kernels, which can faithfully characterize the similarity between images by mimicking deep neural networks; we introduce anchors to aggregate those deep local match kernels into a global match kernel between image sets, which is learned in a supervised way by kernel alignment and therefore more discriminative. The DMK provides the first match kernel framework for image-set classification, which removes specific assumptions usually required in previous approaches and is computationally more efficient. We conduct extensive experiments on four datasets for three diverse image-set classification tasks. The DMK achieves high performance and consistently surpasses state-of-the-art methods, showing its great effectiveness for image-set classification.

1. Introduction

Image-set classification has been one of the most important tasks in computer vision [15, 25, 33, 20, 19, 22, 36, 9, 34, 37, 52] due to its broad applications in various areas including multi-view visual recognition, video-based surveillance, dynamic scene recognition, etc. In contrast to the conventional tasks on one single image [3, 55], in image-set classification, each sample is a set of images, which, therefore, is able to better describe objects in images since each image contains certain more information of variations of the objects. However, compared to classification on single images, image sets exhibit huge intra-class variability and large inter-class ambiguity, which poses great challenges to faithfully measure the similarity between image sets for accurate classification [20].

Image-set classification has been extensively studied in the previous work, which was mostly developed under specific assumptions on image distributions or geometrical structures. In order to facilitate modeling image sets, some specific assumptions, *e.g.*, a single Gaussian [45], Gaussian mixture models [1, 53], on the distribution of images in a set, were made *a priori* in early work. The traditional metrics, *e.g.*, Kullback-Leibler (KL) divergence, were chosen to measure the similarity between distributions of images in sets. However, these methods would not guarantee satisfactory performance when there is no significant statistical relationship between training and test sets due to the huge intra-class variability [31]. The symmetric positive definite (SPD) matrices [52, 21, 28] have been extensively used to represent image sets by computing the second-order statistic, *e.g.*, the covariance matrix of images in the set. The covariant matrix as a statistic measurement can be too general to handle the heavy inter-class ambiguity due to the lack of local information in each individual image. Those SPDs lie in a specific Riemannian manifold and therefore conventional approaches in the Euclidean space are not directly applicable [52]. In addition, the SPD based representation induces heavy computational cost when the dimensionality of the SPD matrix is high [28]. Another important body of

*Corresponding author.

work are developed under the assumption that image sets lie on a Grassmann manifold [18, 23, 22, 27], where each image set is regarded as a linear subspace on the Grassmann manifold. To measure the similarity between linear subspaces, a family of Grassmann kernels, *e.g.*, the projection and Binet-Cauchy kernels [18, 22] were proposed based on principal angles. However, the principal angle contains only weak information about the location and boundary of the samples in the input space [52], which unfortunately lacks sufficiently discriminative information to deal with the huge intra-class variability.

Although developed in different frameworks, most of the previous approaches essentially manipulate the similarity or distance metric between images from two sets implicitly. In other words, the similarity between image sets is ultimately determined by the similarity between images from the sets. Based on this important observation, in this paper, we propose learning deep match kernels between image sets by directly measuring the similarity a new match kernel framework, which removes prior assumptions on image distributions or geometrical structures while effectively capturing discriminative information localized in each image. Match kernels [35] between image sets involve the local match kernel between images and the global match kernel between sets, which aggregates local match kernels between the pair of images from the sets. In our DMK, local match kernels are built upon the powerful arc-cosine kernel and aggregated into the global match kernel which is learned by kernel alignment via anchors. The framework of constructing the DMK is illustrated in Figure. 1.

For the local match kernel, we propose building a deep local match kernel upon the arc-cosine kernel [10]. Thanks to the nature of mimicking deep neural networks with infinite hidden units, the arc-cosine kernel has the great capability of characterizing the similarity between images. The faithful measurement of similarity between images from two sets by the deep local kernels underpins the construction of global match kernels between image sets.

For the global match kernel, we propose aggregating those deep local match kernels in a supervised way by kernel alignment via anchors, which enables it to conquer inter-class ambiguity and intra-class variability. The anchor-based global match kernel is not only highly discriminative by exploring different discriminative abilities of local match kernels but also computationally more efficient compared to conventional match kernels.

The major contributions of this work can be summarized in the following three aspects:

- We propose the first match kernel framework, deep match kernels (DMK), for image-set classification, which removes specific assumptions on distributions or representations of sets. The DMK can effectively and efficiently characterize the similarity between image sets by directly matching images.
- We build deep local match kernels on the arc-cosine kernel to faithfully measure the similarity between images. The deep local match kernel leverages the nature of arc-cosine kernels to mimic the computation of deep neural networks with an infinite number of units.
- We introduce anchors to aggregate those deep local match kernels into global match kernels between image sets, which are learned by kernel alignment. The anchor-based aggregation provides a new supervised learning framework to establish kernels between image sets by exploring the different discriminative abilities of local match kernels.

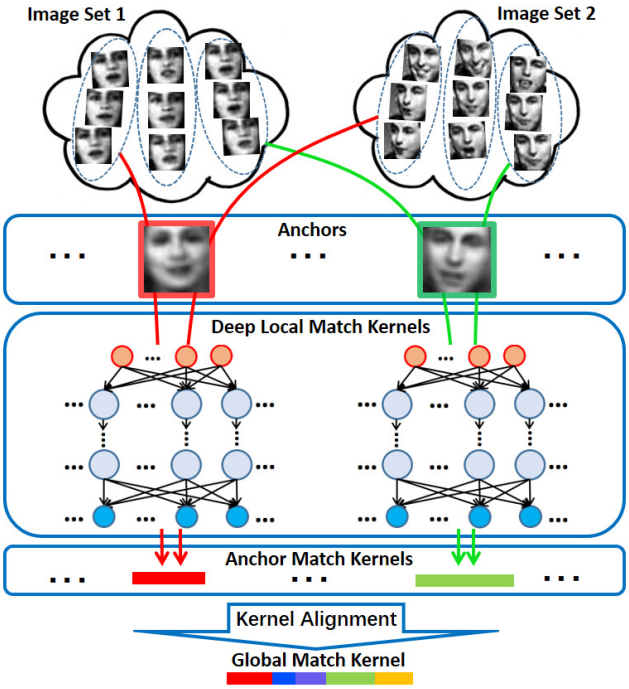

Figure 1. The framework of the deep match kernels (DMK). Images assigned to anchors are matched by deep local match kernels which are summed to anchor match kernels. Global match kernels are obtained by kernel alignment of anchor match kernels.

The DMK has been evaluated by extensive experiments on four datasets for three challenging computer vision tasks, which has produced high performance and consistently surpassed state-of-the-art algorithms.

2. Related work

With the great potential of practical use in widespread applications, image-set classification has been widely studied in the last few decades [45, 31, 18]. Due to the great inter-class ambiguity and high inter-class variability, it is challenging to measure the similarity between image sets,

which usually contain different cardinalities of images. Previous work has been developed under certain specific assumptions on the distribution of images in sets or on the geometrical structures of the data.

Due to the capability of characterizing the distributions of image sets, statistical models have been explored to model image sets in early work [45, 1]. Under the prior assumptions of Gaussian distributions, single multivariate Gaussian model [45] and Gaussian mixture models [1] were used to represent image sets. The widely used metrics, *e.g*., Kullback-Leibler (KL) divergence, were chosen to measure the similarity between distributions. However, usually sufficient samples are required to well estimate the parameters of the distributions and those models would not perform well when there is no strong statistical correlation between training and testing data [31, 53].

Symmetric positive definite (SPD) matrices [52] are proposed to model an image set with its second-order statistic, *e.g*., the covariance matrix. SPDs are assumed to lie on Riemannian manifolds, and the Log-Euclidean distance [2] that projects a point from the Riemannian manifold to the Euclidean space is chosen to measure the distance between SPDs. Although it is natural to characterize a set structure using the SPD matrix, it tends to be computationally very expensive due to the high dimensionality of the SPDs. Moreover, as indicated in [21], there would unavoidably induce distortions during the flattening from Riemannian manifold to the Euclidean space. To overcome these limitations, Harandi *et al*. [21] model the mapping from high-dimensional SPD manifold to a low-dimensional one with an orthogonal projection. Similarly, Huang *et al*. [28] propose Log-Euclidean metric learning to directly map an original tangent space to a more discriminative tangent space. Lu *et al*. [34] extend the second-order statistic to multiple order statistics including the mean vector, the covariance matrix and the Kronecker product between them. Unfortunately, the obtained features of images can be in the third order of magnitude of the original feature vectors, which will induce very high computational cost.

The Grassmann manifold has been playing an important role in image-set classification [18]. The assumption of methods based on Grassmann manifolds is that a set of images can be well approximated by a low dimensional subspace. Then, discriminant analysis methods are introduced on Grassmann manifolds [27, 19, 23]. Kernel methods [44] showing great effectiveness for both classification [54] and regression [56], have also been explored for image-set classification. A family of positive definite kernels on the Grassmann manifold of image sets is developed [14, 50, 18, 22], which indicates the great potential of directly matching image sets by kernels.

To handle the large variations of image appearance in the set, affine hull or convex hull models [6, 26] have been introduced to model image sets. The distances between image sets are measured by geometric distances between convex models or sparse approximated nearest points (SANP). However, due to the affine/linear subspace assumption, they would not able to handle the highly nonlinear variations of image appearance, and moreover, the performance is prone to outliers because of the used inter-point distance [25]. In addition, the computational cost can be too expensive due to the requirement of the one-to-one match for a query set.

Most of the above methods were developed under certain specific assumptions, which would not hold in practice or be shared across different applications. We propose deep match kernels (DMK) under the match kernel framework [35, 16], which removes those assumptions and provides a direct measurement between image sets.

3. Deep Match Kernels

The major challenge in image-set classification is to faithfully measure the similarity between sets. We propose directly learning the similarity by deep match kernels (DMK) in the match kernel framework. The DMK builds local match kernels on the arc-cosine kernel which mimics a deep infinite neural network with the strong capability of measuring the similarity of images; these local match kernels are aggregated via anchors into a global match kernel between image sets, which is learned by kernel alignment.

3.1. Preliminaries

We briefly revisit two fundamental concepts, *e.g*., the kernel between sets and match kernels. We reveal that the kernel between two image sets is essentially characterized by the similarity between images from the two sets, which motivates us to learn the kernel between image sets by directly matching images from them.

3.1.1 Kernel between Sets

We start with the distance between sets of vectors, *e.g*., image sets, which will be used to construct the kernel between sets in image-set classification. To keep general, we consider the distance between image distributions of sets.

Given two image sets a and b denoted by $X_a = \{\mathbf{x}_i^{(a)}\}_{i=1}^{|X_a|}$ and $X_b = \{\mathbf{x}_i^{(b)}\}_{i=1}^{|X_b|}$, respectively, without loss of generality, the distance between their distributions $p_a(\mathbf{x})$ and $p_b(\mathbf{x})$ can be measured by the Hellinger distance [5, 20] as follows:

$$D_H^2(p_a \| p_b) = \frac{1}{|X_a|} \sum_{i=1}^{|X_a|} \left(\sqrt{R(\mathbf{x}_i^{(a)})} - \sqrt{1 - R(\mathbf{x}_i^{(a)})} \right)^2 + \frac{1}{|X_b|} \sum_{i=1}^{|X_b|} \left(\sqrt{R(\mathbf{x}_i^{(b)})} - \sqrt{1 - R(\mathbf{x}_i^{(b)})} \right)^2, \quad (1)$$

where

$$R(\mathbf{x}) = \frac{p_a(\mathbf{x})}{p_a(\mathbf{x}) + p_b(\mathbf{x})}.$$

The distance is indeed a function of $R(\mathbf{x})$. By using kernel density estimators [39] with a bandwidth of h, we obtain

$$R(\mathbf{x}) = \frac{\frac{1}{|X_a|}\sum_{i=1}^{|X_a|} k(\frac{\mathbf{x}-\mathbf{x}_i^{(a)}}{h})}{\frac{1}{|X_a|}\sum_{i=1}^{|X_a|} k(\frac{\mathbf{x}-\mathbf{x}_i^{(a)}}{h}) + \frac{1}{|X_b|}\sum_{i=1}^{|X_b|} k(\frac{\mathbf{x}-\mathbf{x}_i^{(b)}}{h})}. \tag{2}$$

We can observe from (1) and (2) is that the kernel between image sets can be ultimately calculated by measuring the similarity/distance between each pair of images from them, which indicates that we can directly find a match kernel between two image sets by matching images from the two sets. We revisit the match kernel framework in Sec. 3.1.2, which serves as the theoretical foundation for the derivation of our deep match kernels (DMK).

3.1.2 Match Kernels

Match kernels as fundamental tools have been widely used in computer vision and machine learning [35, 16, 4], which provide a direct effective way to measure the similarity between two sets of feature vectors, *e.g.*, image sets. A widely used match kernel between two sets of feature vectors is the sum match kernel defined as follows:

Definition 1 (Sum Match Kernel [35])**.** *Let $X_a = \{\mathbf{x}_i^{(a)}\}_{i=1}^{|X_a|}$ and $X_b = \{\mathbf{x}_i^{(b)}\}_{i=1}^{|X_b|}$ be two image sets, the normalized summation of match kernel is defined as:*

$$K(X_a, X_b) = \frac{1}{|X_a|}\frac{1}{|X_b|}\sum_{i=1}^{|X_a|}\sum_{j=1}^{|X_b|} k(\mathbf{x}_i^{(a)}, \mathbf{x}_j^{(b)}), \tag{3}$$

where $k(\mathbf{x}_i^{(a)}, \mathbf{x}_j^{(b)})$ is the local match kernel between features vectors $\mathbf{x}^{(a)}$ and $\mathbf{x}^{(b)}$ from X_a and X_b respectively.

A new Mercer kernel was introduced in [35] by replacing the local match kernel in (3) with $\left[k(\mathbf{x}_i^{(a)}, \mathbf{x}_j^{(b)})\right]^p$, where $p \geq 1$. To guarantee the convergence of learning algorithms and existence of a unique global optimal solution, match kernels are required to satisfy the Mercer condition [46]. We introduce the definition of Mercer kernels and their closure properties, which will be used to construct our DMK.

Definition 2 (Mercer Kernel [46])**.** *Let $\mathcal{X}$ be any input space and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a symmetric function, k is a Mercer kernel, also known as positive definite kernel, if and only if the kernel matrix formed by restricting k to any finite subset of $\mathcal{X}$ is positive definite.*

The following closure properties of the positive definite kernels are widely adopted to construct Mercer kernels.

1. If two kernels k_1 and k_2 are positive definite (p.d.), then so is their linear combination $a_1k_1 + a_2k_2$, where $a_1, a_2 \geq 0$. [42]
2. Let k be a p.d. kernel defined on $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, for any finite $A, B \subseteq \mathcal{X}$, define $k'(A, B) = \sum_{\mathbf{x}\in A}\sum_{\hat{\mathbf{x}}\in B} k(\mathbf{x}, \hat{\mathbf{x}})$. Then k' is a p.d. kernel. (Lemma 1 in [24])

When constructing global match kernels between image sets, the following properties are highly desired.

- The local match kernel should faithfully reflect the similarity between images.
- The global match kernel should satisfy the Mercer condition, *e.g.*, positive definitiveness.
- The different discriminative ability of local match kernels should be distinguished when aggregating into global match kernel.
- The computation of global match kernels should be efficient in both time and space.

We propose deep match kernels which simultaneously address the above issues and achieve discriminative and computationally efficient kernels between image sets.

3.2. Deep Local Match Kernel

We propose building deep local match kernels upon the arc-cosine kernel [10] to leverage its great capability of measuring the similarity between images. By mimicking the computation in deep learning networks of infinite units, the arc-cosine kernel outperforms the widely used radius basis function (RBF) kernel [43], which can be viewed as a single-layer infinite network [40].

Specifically, the r-th order arc-cosine kernel between two vectors, $\mathbf{x}, \hat{\mathbf{x}} \in \mathbb{R}^{\mathrm{d}}$, is defined in an integral representation as follows:

$$k_r(\mathbf{x}, \hat{\mathbf{x}}) = 2\int \frac{e^{-\frac{\|\mathbf{w}\|^2}{2}}}{(2\pi)^{d/2}}\Theta(\mathbf{w}\cdot\mathbf{x})\Theta(\mathbf{w}\cdot\hat{\mathbf{x}})(\mathbf{w}\cdot\mathbf{x})^r(\mathbf{w}\cdot\hat{\mathbf{x}})^r d\mathbf{w}, \tag{4}$$

where $\Theta(z) = \frac{1}{2}(1 + \mathrm{sign}(z))$ denotes the Heaviside step function. (4) can be viewed as the dot product computation between infinite dimensional outputs of a single-layer neural network with Gaussian random weights $\mathbf{w}$ and the activation function

$$g_r(z) = \Theta(z)z^r. \tag{5}$$

The arc-cosine kernel is highly flexible in that $g_r(\cdot)$ can achieve the step function, ramp function with rectification nonlinearity [17] and the quarter-pipe function by setting $r = 0, 1, 2$, respectively. With different orders r, the activation function $g_r(z)$ has different abilities of handling nonlinearity, which significantly increases the capability of the representation in neural networks.

The arc-cosine kernel in (4) can be analytically computed [10] by

$$k_r(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{\pi}\|\mathbf{x}\|^r \|\hat{\mathbf{x}}\|^r J_r(\theta), \tag{6}$$

which is composed of the magnitudes of input vectors and the angle between them. The angular dependence function $J_r(\theta)$ is defined as

$$J_r(\theta) = (-1)^r (\sin\theta)^{2r+1} \Big(\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\Big)^r \Big(\frac{\pi-\theta}{\sin\theta}\Big) \tag{7}$$

where $\theta = \arccos\Big(\frac{\langle \mathbf{x}, \hat{\mathbf{x}}\rangle}{\|\mathbf{x}\|\|\hat{\mathbf{x}}\|}\Big)$.

We provide the formulations of the first four orders, $r = 0, 1, 2, 3$, of the angular dependence function $J_r(\theta)$, which will be used in our deep match kernels.

$$\begin{aligned}
J_0(\theta) &= \pi - \theta, \\
J_1(\theta) &= (\pi - \theta)\cos\theta + \sin\theta, \\
J_2(\theta) &= (\pi - \theta)(1 + 2\cos^2\theta) + 3\sin\theta\cos\theta, \\
J_3(\theta) &= (\pi - \theta)(9\sin^2\theta\cos\theta + 15\cos^3\theta) + 4\sin^3\theta \\
&\quad + 15\sin\theta\cos^2\theta.
\end{aligned}$$

By the arc-cosine kernels, the inputs $\mathbf{x}$ and $\hat{\mathbf{x}}$ are matched by transforming them through the infinite network with an activation function $g_r(\cdot)$, which achieves a deep local match kernel to measure the similarity between images.

The kernel function can be viewed as inducing a nonlinear mapping from inputs $\mathbf{x}$ to a high even infinite dimensional feature vector $\phi(\mathbf{x})$. The power of the arc-cosine kernel stems from its ability to achieve deep learning with multiple layers by applying ℓ successive times of nonlinear mapping $\phi(\cdot)$.

$$k^{(\ell)}(\mathbf{x}, \hat{\mathbf{x}}) = \langle \underbrace{\phi(\phi(\cdots\phi(\mathbf{x})))}_{\ell \text{ times}}, \underbrace{\phi(\phi(\cdots\phi(\hat{\mathbf{x}})))}_{\ell \text{ times}} \rangle. \tag{8}$$

This can be computed efficiently due to the nested compositions of kernels rather than explicitly training a multi-layer neural network [10]. Specifically, the construction of the arc-cosine kernels for ℓ-layer networks is given by

$$k_r^{(\ell+1)}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{\pi}\big[k_r^{(\ell)}(\mathbf{x}, \mathbf{x}) k_r^{(\ell)}(\hat{\mathbf{x}}, \hat{\mathbf{x}})\big]^{\frac{r}{2}} J_r(\theta_r^{(\ell)}). \tag{9}$$

where $\theta_r^{(\ell)}$ is the angle between the inputs in feature space induced by ℓ-fold composition and can be written as

$$\theta_r^{(\ell)} = \arccos\left(\frac{k_r^{(\ell)}(\mathbf{x}, \hat{\mathbf{x}})}{\sqrt{k_r^{(\ell)}(\mathbf{x}, \mathbf{x}) k_r^{(\ell)}(\hat{\mathbf{x}}, \hat{\mathbf{x}})}}\right). \tag{10}$$

The obtained local match kernels essentially accomplish deep learning to construct kernels between images, which we, therefore, refer as deep local match kernels.

3.3. Anchor Global Match Kernel

We propose aggregating those deep local match kernels into a global kernel between image sets by introducing anchors based on which kernel alignment is employed to learn discriminative kernels between image sets.

3.3.1 Anchor Match Kernel

Introducing anchors for aggregating local match kernels brings us two desirable benefits: 1) we are able to explore the different discriminant abilities of local match kernels by learning the weights associated with anchors in a supervised way; 2) and we are able to compute more efficiently by just matching images assigned to the same anchors.

We first construct a set of M anchors $C = \{\mathbf{c}_m\}_{m=1}^M$ by quantizing all the images from the training samples by the k-means clustering algorithm. Images from each set are then assigned to anchors. Unlike traditional assignment methods, for each anchor, we find the n nearest images from each image sets and assign them to this anchor, which avoids empty anchors.

We then compute the match kernel K_m between images assigned to each anchor $\mathbf{c}_m \in C$, where K_m is referred as the anchor match kernel. Specifically, the anchor match kernel is the sum of local match kernels between images assigned to anchors. Specifically, given two sets $X_a = \{\mathbf{x}_i^{(a)}\}_{i=1}^{|X_a|}$ and $X_b = \{\mathbf{x}_i^{(b)}\}_{i=1}^{|X_b|}$, the anchor match kernel is defined as:

$$\begin{aligned}
&K_m(X_a, X_b) = \\
&\frac{1}{n^2} \sum_{\mathbf{x}_i^{(a)} \in N_{n,\mathbf{c}_m}^{(a)}} \sum_{\mathbf{x}_j^{(b)} \in N_{n,\mathbf{c}_m}^{(b)}} k(\mathbf{c}_m - \mathbf{x}_i^{(a)}, \mathbf{c}_m - \mathbf{x}_j^{(b)})
\end{aligned} \tag{11}$$

where $N_{n,\mathbf{c}_m}^{(a)}$ denotes the n nearest neighbors of the m-th anchor in the set a, and $N_{n,\mathbf{c}_m}^{(b)}$ is defined similarly. The use of the difference between images and anchors is inspired by its success in the construction of vector of locally aggregated descriptors (VLAD) [29], which has shown great effectiveness in image representations.

Having anchor match kernels in (11), the global match kernel between the two image sets is obtained by

$$K_G(X_a, X_b) = \sum_{m=1}^M \omega_m K_m(X_a, X_b) \tag{12}$$

where $\boldsymbol{\omega} = \{\omega_1, \cdots, \omega_m, \cdots, \omega_M\}$ with $\boldsymbol{\omega} \geq \mathbf{0}$ are the weight coefficients associated with anchor match kernels.

The positive definiteness of the obtained global match kernel is crucial to the robust solution with a unique optimum which is guaranteed by Theorem 1.

Theorem 1 (Positive Definiteness of Anchor Global Match Kernel)**.** *The anchor global match kernel in (12) satisfies the Mercer condition and is, therefore, positive definite.*

The positive definiteness of the anchor global match kernel is essentially guaranteed by deep local match kernels built on arc-cosine kernels. Indeed, the arc-cosine kernel can be viewed as the inner product between high-dimensional feature maps from the neural network of infinite units. Denote $\{\mathbf{w}_i\}_{i=1}^h$ as ith row of the weight matrix W of the network with the activation function in (5). The inner product is

$$g_r(W\mathbf{x})\cdot g_r(W\hat{\mathbf{x}}) = \sum_{i=1}^{h} \Theta(\mathbf{w}_i\cdot\mathbf{x})\Theta(\mathbf{w}_i\cdot\hat{\mathbf{x}})(\mathbf{w}_i\cdot\mathbf{x})^r(\mathbf{w}_i\cdot\hat{\mathbf{x}})^r. \tag{13}$$

which induces a positive definite kernel. The arc-cosine kernel can be obtained with $h \to \infty$, *e.g.*,

$$k_r(\mathbf{x},\hat{\mathbf{x}}) = \lim_{h\to\infty} g_r(W\mathbf{x})\cdot g_r(W\hat{\mathbf{x}}). \tag{14}$$

Therefore, the arc-cosine kernel is positive definite. Since the anchor global match kernel is aggregated from deep local match kernels built on arc-cosine kernels by a linear combination with $\boldsymbol{\omega} \geq \mathbf{0}$, therefore we can straightforwardly derive the positive definiteness from closure properties $1, 2$ under the definition of Mercer kernels (Definition 2).

3.3.2 Learning by Kernel Alignment

We propose learning the weight coefficients $\boldsymbol{\omega}$ associated with anchors in a supervised way by kernel target alignment, which has shown great effectiveness in learning the optimal combination of multiple kernels [12, 11].

The core idea of kernel alignment is to align an input kernel K to a target kernel K_T by maximizing the similarity or the degree of agreement between them. Specifically, the alignment between kernels is defined as

$$A(K, K_T) = \frac{\langle K, K_T\rangle_F}{\sqrt{\langle K, K_T\rangle_F\langle K_T, K_T\rangle_F}}. \tag{15}$$

Intuitively, the measurement of alignment can be viewed as the cosine of the angle between two bi-dimensional vectors K and K_T. Kernel alignment offers a best-suited way to obtain the weight coefficients $\boldsymbol{\omega}$. We now introduce the kernel alignment formulation to learn our anchor global match kernel. We would like to maximize the alignment between the target kernel matrix K_T and the global kernel $K_G(\boldsymbol{\omega})$ denoted as $K_{\boldsymbol{\omega}}$ for simplicity, and based on (15), we have the following optimization problem

$$\boldsymbol{\omega}^* = \arg\max A(K_{\boldsymbol{\omega}}, K_T) = \arg\max \frac{\mathrm{Tr}(K_{\boldsymbol{\omega}}K_T)}{\sqrt{\mathrm{Tr}(K_{\boldsymbol{\omega}}K_{\boldsymbol{\omega}})}}. \tag{16}$$

The target kernel matrix K_T is constructed by defining the target kernel $K_T = YY^\top$, where Y is the matrix composed of the class label vectors, *e.g.*, $Y = [\mathbf{y}_1, \cdots, \mathbf{y}_i, \cdots, \mathbf{y}_N]^\top$ for N samples, and $\mathbf{y}_i$ is a binary vector of the length of classes, in which only the c-th element is 1 if $\mathbf{x}_i$ is from the c-th class.

As indicated in [11], to obtain the high correlation between performance and kernel alignment, it is necessary to center all kernel matrices K_m before alignment. Let $[K_m]_{ij}$ denote the element in K_m and the centered kernel matrix can be computed by

$$\begin{aligned}\big[\bar{K}_m\big]_{ij} =& \big[K_m\big]_{ij} - \frac{1}{N}\sum_{i=1}^{N}\big[K_m\big]_{ij} \\ &- \frac{1}{N}\sum_{j=1}^{N}\big[K_m\big]_{ij} + \frac{1}{N^2}\sum_{i,j=1}^{N}\big[K_m\big]_{ij}.\end{aligned} \tag{17}$$

We can further equivalently rewrite the objective function in (16) as follows:

$$\boldsymbol{\omega}^* = \underset{\|\boldsymbol{\omega}\|=1,\boldsymbol{\omega}\geq 0}{\arg\max} \frac{\boldsymbol{\omega}^\top\boldsymbol{\beta}\boldsymbol{\beta}^\top\boldsymbol{\omega}}{\boldsymbol{\omega}^\top\Omega\boldsymbol{\omega}} \tag{18}$$

where $\boldsymbol{\omega} \geq 0$ guarantees the positive definiteness, $\|\boldsymbol{\omega}\| = 1$ is a regularization term, for $i, j \in \{1, \cdots, M\}$, $\boldsymbol{\beta}$ is defined by $\boldsymbol{\beta}_i = \mathrm{Tr}\left(\bar{K}_i K_T\right)$ and the matrix Ω is defined by $\Omega_{ij} = \mathrm{Tr}\left(\bar{K}_i\bar{K}_j\right)$.

This alignment maximum problem in (18) can be reduced to a simple quadratic programming (QP) problem [38] as shown in the Proposition 1, which does not require the inversion of Ω in (18) and can be solved efficiently.

Proposition 1. *Let* $\mathbf{q}^*$ *be the solution of the following QP:*

$$\mathbf{q}^* = \underset{\mathbf{q}\geq 0}{\arg\min}\, \mathbf{q}^\top\Omega\mathbf{q} - 2\mathbf{q}^\top\boldsymbol{\beta}. \tag{19}$$

Then, the solution $\boldsymbol{\omega}^*$ *of the alignment maximization problem (18) is given by*

$$\boldsymbol{\omega}^* = \frac{\mathbf{q}^*}{\|\mathbf{q}^*\|}. \tag{20}$$

Proof. The proof can be referred to the proof of Proposition 3 in [11]. □

3.3.3 Complexity Analysis

Due to the introduced anchors, the proposed deep match kernels (DMK) are computationally more efficient than conventional match kernels, *e.g.*, the sum match kernel (SMK). We provide the time complexity analysis to show the efficiency advantage. The complexity of match kernels is mainly induced by the computation of the kernel between two image sets. Given N image sets with a maximum of L images, M anchors, and n nearest neighbors,

the time complexity of our DMK is $\mathcal{O}(N^2n^2M)$ in comparison to $\mathcal{O}(N^2L^2)$ for the SMK. Given a typical setting with $M = 50$, $n = 10$ and $L = 500$, $n^2M(= 5,000) \ll L^2(= 250,000)$. Therefore, the time complexity is largely reduced in our DMK compared to the SMK.

4. Experiments

We show the effectiveness of the proposed deep match kernels (DMK) on three challenging computer vision tasks, *e.g.*, video-based face recognition, dynamic scene classification, and set-based object categorization.

4.1. Experimental Settings

We set the parameters ℓ, r in arc-cosine kernels to be $\ell = 4$, $r = [0, 1, 3, 3]$, respectively by cross-validation, which generally produces the best overall performance on all the datasets. The number of anchors and neighbors are the key parameters to be set, which have been thoroughly investigated on all the datasets in our experiments. We implement two variants of sum match kernels (SMK) (3) as baseline match kernels. We adopt the support vector machine (SVM) [8] with the setting of pre-computed kernels for classification.

We compare with representative state-of-the-art algorithms including subspace-based modeling for DCC [31], GDA [18], MDA [51], PML [51], GEDA [23], CDL [52], SPD-ML [21], LEML [28], DARG [53] and MPDF [20]. Specifically, we implement 3 variants of MPDF, namely kFDA-J, kFDA-HL and NN-J-DR. The default parameters of all these methods are tuned by following the original work. For DCC, PCA is performed to learn the subspace by keeping 90% energy. The numbers of basis vectors for subspace in GDA, MDA and GEDA are chosen by cross-validation and we report the best result. The parameter of dimension d in PML [51] is chosen as reported by the author. LDA is used for discriminative learning in CDL [52]. For LEML, two parameters η and ζ are searched in the range of $[0.1, 1, 10]$ and $[0.1 : 0.1 : 1]$ respectively. For DARG, the number of Gaussian components in GMM is set to 7 as suggested by the authors [53].

4.2. Results

The proposed DMK consistently produces the high performance on all tasks and largely outperforms the baseline sum match kernels (SMK) and representative state-of-the-art algorithms. The results are reported in Tables 1. In what follows, we provide the implementation and comparison details on each task.

4.2.1 Video-based Face Recognition

We conduct experiments for video-based face recognition on the commonly used YTC dataset [30] which contains 1910 video clips of 47 subjects. This dataset exhibits large diversity in terms of illumination, facial expressions, and poses. There are hundreds of frames in each clip. By following settings in previous work [25], we adopt the algorithm in [41] to detect the faces for each clip and resize to patches of the size 50×50. The local binary pattern (LBP) [49] is used for face description, which is reduced to 1000 by PCA.

For the fair comparison, we follow the standard validation protocol [34]; specifically, for each subject, we randomly choose 9 videos with 3 and 6 for training and query sets, respectively. The results are the average from five times. We set parameters M, the number of anchors, to be 100 and n, the number of nearest neighbors to be 4, respectively. As shown in Table 1 (3^{rd} column), our DMK achieves the highest identification rate of 80.3%.

4.2.2 Dynamic Scene Classification

Dynamic scene classification has been an important task in computer vision, which has recently been addressed as image-set classification. We show the advantage of our DMK on two datasets, *e.g.*, the UCSD [7] and MDSD [47] datasets for this task. For UCSD, we compute the HoG features [13] to describe each frame in videos. We follow the training/test split settings shared in [20]. The parameters for this dataset are set as $M = 10$ and $n = 3$. For the MDSD, with 10 videos per class, the dataset contains 13 different classes of dynamic scenes. The task is very challenging because scenes in the wild are unconstrained with large variation in scale, view, illumination, background. We choose the last fully connected layer of the CNN [57, 48] as the descriptor for each frame and reduce the dimensionality of the CNN features from 1183 to 400 by PCA.

Following the settings in [20], we test the method based on two protocols, *e.g.*, standard leave-one-out (LOO) and seventy-thirty-ratio (STR) which partitions the dataset into gallery and probes by randomly choosing 7 videos for training and 3 videos for testing in each class. The parameters in MDSD dataset are set as $M = 10$ and $n = 100$. As shown in Table 1 (4^{th}-6^{th}columns), on the two datasets, our DMK surpasses all the compared methods.

4.2.3 Set-based Object Categorization

Set-based object classification is an important computer vision task. We experiment on the ETH-80 dataset [32], which has been widely used for set-based object classification. There are 41 images for each set of different orientations. To achieve the fair comparison with other methods, we follow the same experimental setup in [52, 34, 33]. Each image is segmented from all the simple background and scaled 20×20 for classification. For each object, 5 instances are selected as the gallery and the remaining five are

Table 1. The performance comparison on the YTC, UCSD, MDSD, ETH-80 datasets.

Method \ Dataset	Years	YTC	UCSD	MDSD-STR	MDSD-LOO	ETH-80
DCC [31]	2007	65.4 ± 3.9	91.5 ± 3.4	69.8 ± 6.1	80.5 ± 5.5	91.7 ± 9.0
GDA [18]	2008	66.0 ± 6.9	92.5 ± 2.6	70.4 ± 4.5	81.5 ± 5.1	95.0 ± 3.9
MDA [51]	2009	67.2 ± 4.0	92.7 ± 3.6	72.3 ± 4.2	82.4 ± 3.0	89.0 ± 2.0
GEDA [23]	2011	69.3 ± 2.2	92.4 ± 2.3	70.3 ± 5.2	82.2 ± 6.1	92.3 ± 2.4
CDL [52]	2012	70.1 ± 4.6	91.7 ± 0.9	76.7 ± 7.8	86.5 ± 5.8	91.5 ± 3.5
SPD-ML [21]	2014	69.8 ± 6.7	92.1 ± 1.5	77.3 ± 6.2	84.3 ± 7.2	93.2 ± 5.3
PML [27]	2015	70.3 ± 3.7	94.7 ± 3.1	72.4 ± 3.7	82.7 ± 3.7	95.5 ± 4.3
LEML [28]	2015	73.3 ± 2.9	92.5 ± 2.9	77.6 ± 5.2	86.5 ± 6.2	96.0 ± 2.1
DARG [53]	2015	77.1 ± 4.3	95.5 ± 3.0	73.6 ± 4.4	83.5 ± 5.8	92.3 ± 2.4
kFDA-J [20]	2015	79.3 ± 3.6	97.3 ± 1.4	77.8 ± 5.3	86.9 ± 4.3	93.7 ± 1.4
kFDA-HL [20]	2015	77.5 ± 3.8	96.5 ± 1.5	79.0 ± 3.1	87.1 ± 5.3	93.1 ± 2.0
NN-J-DR [20]	2015	78.1 ± 1.9	95.6 ± 1.5	80.2 ± 3.7	82.3 ± 3.9	93.8 ± 2.8
SMK* $(p = 1)$ [35]		77.5 ± 3.8	97.0 ± 1.3	79.5 ± 3.9	85.7 ± 4.1	93.0 ± 2.9
SMK* $(p = 3)$ [35]		78.1 ± 1.9	97.6 ± 2.4	78.4 ± 4.1	85.9 ± 5.2	93.7 ± 3.8
DMK (Ours)		$\mathbf{80.3 \pm 4.7}$	$\mathbf{98.0 \pm 0.9}$	$\mathbf{81.5 \pm 4.7}$	$\mathbf{87.2 \pm 5.0}$	$\mathbf{96.8 \pm 1.5}$

* p denotes the power of local match kernels in (3).

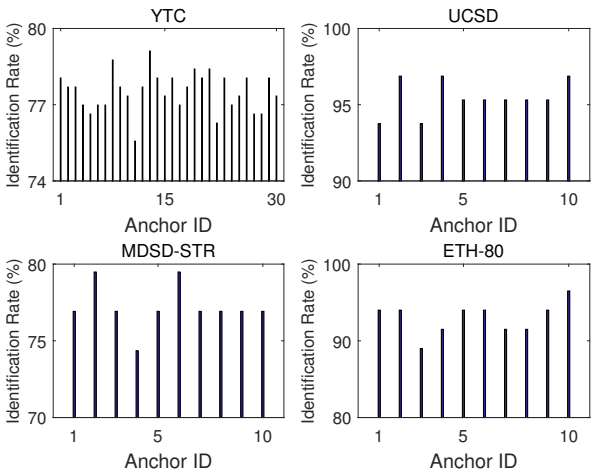

Figure 2. Different discriminant abilities of anchor match kernels.

used for probes. We run 10 times of experiments and perform different random selections of the gallery and probes sets. The parameters are set as $M = 5$ and $n = 30$. As shown in Table 1 (rightmost column), our DMK achieves the highest identification rate of 96.8% which is much better than most of the compared methods.

4.3. Parameter Analysis

The introduced anchors underpin the aggregation of the deep local match kernels into a global match kernel. We provide a comprehensive investigation into the effects of anchors on the performance. We experiment to look into discriminative abilities of anchor match kernels. As shown in Figure 2, match kernels associated with individual anchors produce distinctive identification rates. This indicates that local match kernels carry different discriminative information, which has been explored in our deep match kernel framework compared to the SMK. The results validate the effectiveness of the introduced anchors.

5. Conclusion

In this paper, we have presented the first match kernel framework, the deep match kernel (DMK), for image-set classification, which removes specific assumptions on image distributions and geometrical structures. We build the local match kernels by the arc-cosine kernel to leverage its nature of mimicking deep learning architectures. We introduce anchors to establish a global match kernel between sets, which is learned by kernel alignment. The obtained global match kernel is more discriminative and efficient to compute compared to conventional match kernels. Experiments on four datasets for three challenging computer vision tasks demonstrate that our DMK consistently surpasses state of the arts.

Acknowledgements

This work was supported by Natural Science Foundation of China (Grant No. 61573219, 61571147, and 61472226), NSFC Joint Fund with Guangdong under Key Project No. U1201258, and the Fostering Project of Dominant Discipline and Talent Team of Shandong Province Higher Education Institutions.

References

[1] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell. Face recognition with image sets using manifold density divergence. In *CVPR*, volume 1, pages 581–588. IEEE, 2005.

[2] 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.

[3] A. Berg, J. Deng, and L. Fei-Fei. Large scale visual recognition challenge 2010, 2010. *URL http://www. image-net. org/challenges/LSVRC/2010/index*, 2011.

[4] L. Bo and C. Sminchisescu. Efficient match kernel between sets of features for visual recognition. In *NIPS*, pages 135–143, 2009.

[5] K. M. Carter. *Dimensionality reduction on statistical manifolds*. ProQuest, 2009.

[6] H. Cevikalp and B. Triggs. Face recognition based on image sets. In *CVPR*, pages 2567–2573. IEEE, 2010.

[7] A. B. Chan and N. Vasconcelos. Probabilistic kernels for the classification of auto-regressive visual processes. In *CVPR*, volume 1, pages 846–851. IEEE, 2005.

[8] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[9] L. Chen. Dual linear regression based classification for face cluster recognition. In *CVPR*, pages 2673–2680. IEEE, 2014.

[10] Y. Cho and L. K. Saul. Kernel methods for deep learning. In *NIPS*, pages 342–350, 2009.

[11] C. Cortes, M. Mohri, and A. Rostamizadeh. Two-stage learning kernel algorithms. In *ICML*, pages 239–246, 2010.

[12] N. Cristianini, J. Shawe-Taylor, A. Elisseeff, and J. S. Kandola. On kernel-target alignment. In *NIPS*, pages 367–373, 2002.

[13] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, volume 1, pages 886–893. IEEE, 2005.

[14] 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.

[15] Q. Feng, Y. Zhou, and R. Lan. Pairwise linear regression classification for image set retrieval. In *CVPR*, June 2016.

[16] K. Grauman and T. Darrell. The pyramid match kernel: Efficient learning with sets of features. *JMLR*, 8(Apr):725–760, 2007.

[17] R. H. Hahnloser, H. S. Seung, and J.-J. Slotine. Permitted and forbidden sets in symmetric threshold-linear networks. *Neural computation*, 15(3):621–638, 2003.

[18] J. Hamm and D. D. Lee. Grassmann discriminant analysis: a unifying view on subspace-based learning. In *ICML*, pages 376–383. ACM, 2008.

[19] 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.

[20] M. Harandi, M. Salzmann, and M. Baktashmotlagh. Beyond gauss: Image-set matching on the riemannian manifold of pdfs. In *ICCV*, pages 4112–4120, 2015.

[21] M. T. Harandi, M. Salzmann, and R. Hartley. From manifold to manifold: Geometry-aware dimensionality reduction for spd matrices. In *ECCV*, pages 17–32. Springer, 2014.

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

[23] M. T. Harandi, C. Sanderson, S. Shirazi, and B. C. Lovell. Graph embedding discriminant analysis on grassmannian manifolds for improved image set matching. In *CVPR*, pages 2705–2712. IEEE, 2011.

[24] D. Haussler. Convolution kernels on discrete structures. Technical report, Citeseer, 1999.

[25] M. Hayat, M. Bennamoun, and S. An. Deep reconstruction models for image set classification. *IEEE TPAMI*, 37(4):713–727, 2015.

[26] Y. Hu, A. S. Mian, and R. Owens. Face recognition using sparse approximated nearest points between image sets. *IEEE TPAMI*, 34(10):1992–2004, 2012.

[27] Z. Huang, R. Wang, S. Shan, and X. Chen. Projection metric learning on grassmann manifold with application to video based face recognition. In *CVPR*, pages 140–149, 2015.

[28] Z. Huang, R. Wang, S. Shan, X. Li, and X. Chen. Log-euclidean metric learning on symmetric positive definite manifold with application to image set classification. In *ICML*, pages 720–729, 2015.

[29] H. Jegou, F. Perronnin, M. Douze, J. Sánchez, P. Perez, and C. Schmid. Aggregating local image descriptors into compact codes. *IEEE TPAMI*, 34(9):1704–1716, 2012.

[30] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley. Face tracking and recognition with visual constraints in real-world videos. In *CVPR*, pages 1–8. IEEE, 2008.

[31] T.-K. Kim, J. Kittler, and R. Cipolla. Discriminative learning and recognition of image set classes using canonical correlations. *IEEE TPAMI*, 29(6):1005–1018, 2007.

[32] B. Leibe and B. Schiele. Analyzing appearance and contour based methods for object categorization. In *CVPR*, volume 2, pages II–409. IEEE, 2003.

[33] J. Lu, G. Wang, W. Deng, P. Moulin, and J. Zhou. Multi-manifold deep metric learning for image set classification. In *CVPR*, pages 1137–1145, 2015.

[34] J. Lu, G. Wang, and P. Moulin. Image set classification using holistic multiple order statistics features and localized multi-kernel metric learning. In *ICCV*, pages 329–336, 2013.

[35] S. Lyu. Mercer kernels for object recognition with local features. In *CVPR*, volume 2, pages 223–229. IEEE, 2005.

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

[37] 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. *IEEE TIP*, 22(12):5252–5262, 2013.

[38] J. Nocedal and S. Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

[39] E. Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076, 1962.

[40] Q. Que and M. Belkin. Back to the future: Radial basis function networks revisited. In *AISTATS*, pages 1375–1383, 2016.

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

[42] I. Schoenberg. Positive definite functions on spheres. *Duke Math. J*, 1:172, 1988.

[43] B. Scholkopf and A. J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

[44] B. Schölkopf and A. J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[45] G. Shakhnarovich, J. W. Fisher, and T. Darrell. Face recognition from long-term observations. In *ECCV*, pages 851–865. Springer, 2002.

[46] J. Shawe-Taylor and N. Cristianini. *Kernel methods for pattern analysis*. Cambridge university press, 2004.

[47] N. Shroff, P. Turaga, and R. Chellappa. Moving vistas: Exploiting motion for describing scenes. In *CVPR*, pages 1911–1918. IEEE, 2010.

[48] X. Sui, Y. Zheng, B. Wei, H. Bi, J. Wu, X. Pan, Y. Yin, and S. Zhang. Choroid segmentation from optical coherence tomography with graph-edge weights learned from deep convolutional neural networks. *Neurocomputing*, 237:332–341, 2017.

[49] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. `http://www.vlfeat.org/`, 2008.

[50] S. Vishwanathan, A. J. Smola, et al. Binet-cauchy kernels. In *NIPS*, pages 1441–1448, 2004.

[51] R. Wang and X. Chen. Manifold discriminant analysis. In *CVPR*, pages 429–436. IEEE, 2009.

[52] R. Wang, H. Guo, L. S. Davis, and Q. Dai. Covariance discriminative learning: A natural and efficient approach to image set classification. In *CVPR*, pages 2496–2503. IEEE, 2012.

[53] W. Wang, R. Wang, Z. Huang, S. Shan, and X. Chen. Discriminant analysis on riemannian manifold of gaussian distributions for face recognition with image sets. In *CVPR*, pages 2048–2057, 2015.

[54] L. Zhang, X. Zhen, and L. Shao. Learning object-to-class kernels for scene classification. *IEEE TIP*, 23(8):3241–3253, 2014.

[55] X. Zhen, Z. Wang, M. Yu, and S. Li. Supervised descriptor learning for multi-output regression. In *CVPR*, pages 1211–1218, 2015.

[56] X. Zhen, M. Yu, X. He, and S. Li. Multi-target regression via robust low-rank learning. *IEEE T-PAMI*, 2017.

[57] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *NIPS*, pages 487–495, 2014.