Discriminative Log-Euclidean Feature Learning for Sparse Representation-Based Recognition of Faces from Videos

Mohammed E. Fathy **Azadeh Alavi** **Rama Chellappa**
Center for Automation Research, University of Maryland
College Park, MD 20742
{mefathy, azadeh, rama} (at) umiacs.umd.edu

Abstract

With the abundance of video data, the interest in more effective methods for recognizing faces from unconstrained videos has grown. State-of-the-art algorithms for describing an image set use descriptors that are either very high-dimensional and/or sensitive to outliers and image misalignment.

In this paper, we represent image sets as dictionaries of Symmetric Positive Definite (SPD) matrices that are more robust to local deformations and outliers. We then learn a tangent map for transforming the SPD matrix logarithms into a lower-dimensional Log-Euclidean space such that the transformed gallery atoms adhere to a more discriminative subspace structure. A query image set is then classified by first mapping its SPD descriptors into the computed Log-Euclidean tangent space and using the sparse representation over the tangent space to decide a label for the image set. Experiments on three public video datasets show that the proposed method outperforms many state-of-the-art methods.

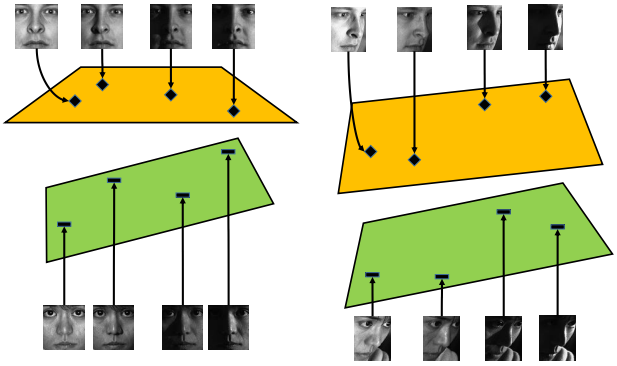

Figure 1: An illustration of the discriminative subspace structure that is naturally exhibited by the *controlled* images of a visual object (e.g. a person's face) [Basri and Jacobs, 2003; Wright *et al.*, 2009]. The example illustrates the property for the face images of two different subjects, taken under two different poses and varying illumination. The images in which the visual object (i.e. face) has the same pose and identity lie close to a low-dimensional subspace regardless of the variations in Lambertian illumination.

1 Introduction

In many practical applications such as surveillance-based face recognition and smartphone video-based face authentication, the test example contains a set of face images that share the same, yet to be determined label. As video-capable consumer devices and surveillance cameras are becoming more abundant, the interest in using image sets for face recognition has grown. While the multiplicity could mean improved recognition, low-resolution and variations in pose, illumination and occlusion limit significant improvements in performance.

Over the years, many methods have been proposed for using image sets for object classification in general and face recognition in particular. In order to capture the variations within an image set and/or model the properties inherent in face images, many methods employ descriptors that live on some non-Euclidean spaces such as the Symmetric Positive-Definite (SPD) manifold or the Grassmann manifold. In such cases, the machine learning algorithms originally designed to work on Euclidean spaces have to be carefully modified to properly and meaningfully work on non-Euclidean ones.

Among the many machine learning algorithms that have been successfully used for image set classification, Sparse Representation-based Classification (SRC) over dictionaries has been shown to be very effective [Chen *et al.*, 2012; Ortiz *et al.*, 2013]. The standard SRC algorithm has become popular in visual identification tasks since the seminal work of Wright *et al.* [2009]. The success of this method is justified by the discriminative low-dimensional subspace structure that is naturally exposed in the space of the visual images of an object. More specifically, it has been mathematically proved that the images of a fixed object taken under varying Lambertian illumination from a fixed viewpoint lie on a low-dimensional subspace [Basri and Jacobs, 2003]. As illustrated in Fig. 1, this suggests that the instances from a particular class lie on (or close to) a low-dimensional linear subspace (assuming same-pose viewing of a static object) or a small number of such subspaces (to account for variations in pose and deformations).

Although originally designed for vector spaces, algorithms for sparse coding have also been extended to work on non-Euclidean spaces such as the SPD manifold [Yuan *et al.*, 2010; Guo *et al.*, 2010]. However, the subspace property was mathematically proved when images are represented using intensities [Basri and Jacobs, 2003]. In other words, the property (mathematically) applies only when pixel raw intensity features or other features derived linearly from the intensities are used. If intensities are nonlinearly transformed (e.g. to extract nonlinear features prior to sparse coding), algorithms based on sparse coding may lose the discriminative advantage offered by the subspace property. Since the SPD descriptors used for image sets are obtained nonlinearly from the input image features, one way to enhance the performance of SRC under that setting is to further embed the nonlinear features in a manner that improves the discriminative subspace structure of the data.

In this paper, we propose an approach for "face identification from image sets" using sparse coding on the Log-Euclidean (LE) tangent space $\mathrm{T}_{\mathbf{I}}\mathbb{S}_+^Q$ of the SPD manifold $\mathbb{S}_+^Q$. In this approach, we first describe each image by generating a number of local covariance descriptors; then we use a dictionary of atoms from the LE tangent space $\mathrm{T}_{\mathbf{I}}\mathbb{S}_+^Q$ to represent each gallery image set. While previous LE approaches for image set classification extract from each image set, a single or very few LE samples that have very high dimensionality, our approach extracts from each set many LE samples of a much lower dimensionality, reducing the possibility of over-fitting. Given the LE features, we then formulate an optimization problem for learning an embedding into a lower-dimensional LE tangent space $\mathrm{T}_{\mathbf{I}}\mathbb{S}_+^q$ in which the data has a more discriminative subspace structure. To classify a probe image set, we use the LE feature transform computed during training to embed the dictionary of LE atoms extracted from the probe image set. Next, we apply the LE sparse coding approach [Yuan *et al.*, 2010; Guo *et al.*, 2010] to classify the embedded probe atoms with respect to the augmented gallery dictionary. Extensive experiments on three public datasets show that our method outperforms many state-of-the-art methods. In addition, we run an empirical ablation analysis to understand how the different components of our approach contribute to its final performance. In order of importance, the contributions of the paper can be summarized as follows:

- An LE dimensionality reduction algorithm that leads to a more discriminative subspace structure, subsequently enhancing the performance of SRC with nonlinear features. Since it reduces the learning problem into that of solving a single generalized eigenvalue in a non-iterative fashion, the algorithm is also efficient.
- An image set feature extractor which models each image set as a dictionary of LE atoms that is more robust to local deformations and has significantly fewer dimensions than other LE image set descriptors [Wang *et al.*, 2012; 2015; Huang *et al.*, 2015b], making our LE features more robust to over-fitting. In our experiments, we show that the proposed features also improve the performance of another recent LE image set method proposed in [Huang *et al.*, 2015b].
- To the best of our knowledge, this paper is the first to apply SRC for image set classification on non-Euclidean spaces. Note that SRC has been extended to image set classification on Euclidean space by Ortiz *et al.* [2013] and to other classification problems on LE tangent spaces [Guo *et al.*, 2010; Yuan *et al.*, 2010]. Our experiments show that the proposed approach outperforms existing methods on three public video face datasets.

The rest of this paper is organized as follows. Section 2 reviews the relevant literature. The proposed approach is presented in Section 3 and the results of the extensive empirical evaluation are presented in Section 4. The paper is concluded in Section 5.

2 Related Work

The image set classification problem has been formulated in various ways. One popular formulation is to compute the distance, either over a vector space or a manifold, between the probe set and each gallery set and then associate the probe with the class of its nearest gallery set. These include discriminative [Hamm and Lee, 2008; Wang and Chen, 2009; Harandi *et al.*, 2011; Wang *et al.*, 2012; Huang *et al.*, 2015b; 2015a; Wang *et al.*, 2015] and non-discriminative methods [Wang *et al.*, 2008; Cevikalp and Triggs, 2010; Hu *et al.*, 2011; Chen *et al.*, 2012; 2013]. Other formulations that do not rely on nearest neighbor-based classification include the binary SVM reverse-training approach [Hayat *et al.*, 2014b], neural network-based methods [Hayat *et al.*, 2014a; Lu *et al.*, 2015], linear representation/coding methods [Ortiz *et al.*, 2013; Zhu *et al.*, 2013] and clustering methods [Mahmood *et al.*, 2014].

Linear Representation (Coding) Methods: The SRC algorithm, originally designed for classification of a single image [Wright *et al.*, 2009], was extended to image sets by Ortiz *et al.* [2013] who proposed the Mean-Sequence SRC (MS-SRC) algorithm. While MS-SRC assumes Euclidean space, Harandi *et al.* extended the sparse coding approach to Grassmann manifold [2013] where it has been applied to face recognition from image sets. Sparse coding over SPD manifolds was also considered but for non-image set classification tasks as in [Yuan *et al.*, 2010; Guo *et al.*, 2010; Harandi *et al.*, 2012]. Since video-based face recognition typically involves large number of gallery samples, we intentionally avoid the scalability issues associated with kernel methods by developing the proposed embedding algorithm and performing sparse coding on the LE tangent space.

Log-Euclidean Feature Learning: Various approaches for learning features, metrics, and/or dimensionality reduction embeddings have been proposed within the LE framework [Wang *et al.*, 2012; Li *et al.*, 2013; Vemulapalli and Jacobs, 2015; Yger and Sugiyama, 2015; Wang *et al.*, 2015; Xu *et al.*, 2015; Huang *et al.*, 2015b]. The goal of these approaches is to boost the performance of nearest neighbor classification whereas the goal of our work is to boost the performance of subspace-based classification. Qiu and Sapiro [2015] proposed an approach for learning linear transformations that improve the performance of subspace-based

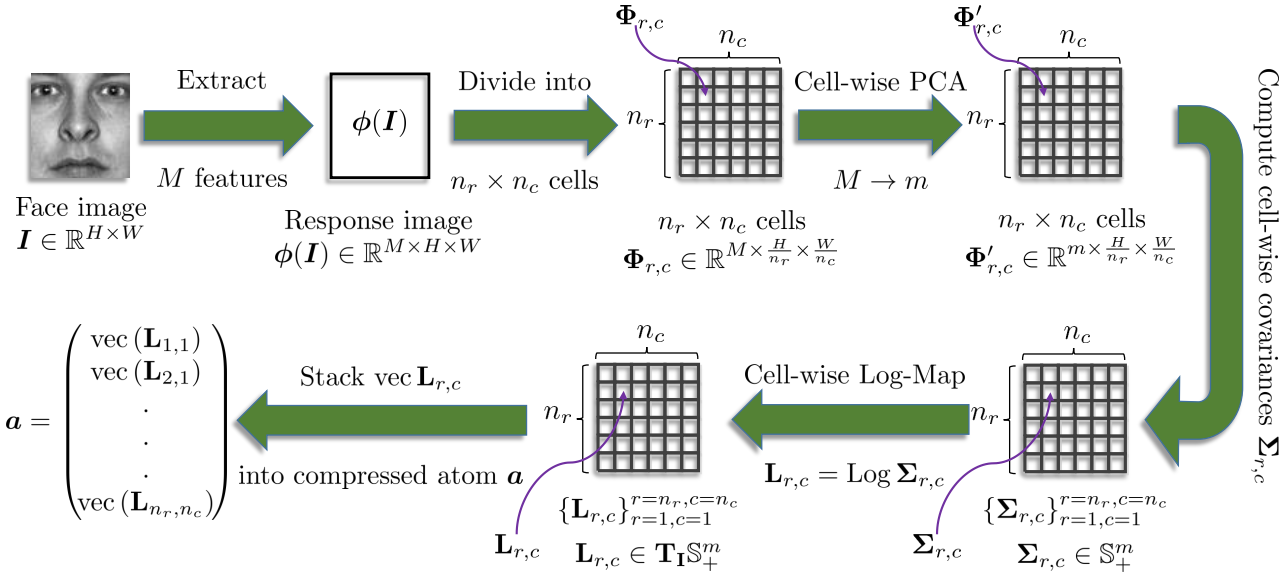

Figure 2: The steps for extracting the LE features from each image.

classification in Euclidean space. The approach uses subgradient descent to minimize a non-convex cost function and so it requires too many iterations to converge and is subject to local minima. In addition, the approach requires performing Singular Value Decomposition (SVD) in each iteration, which makes it even more expensive. The proposed LE feature learning approach is significantly more robust as it is not subject to local minima. In addition, our approach is faster and more scalable as the optimal solution is obtained by solving a single generalized eigenvalue problem.

3 Our Approach

We describe the three components of our approach (description, embedding, and coding) in the following subsections.

3.1 Image Set Descriptor: Dictionary of LE Atoms

Existing SPD image set descriptors, like those used in [Wang *et al.*, 2012; Huang *et al.*, 2015b; Wang *et al.*, 2015] compute a single or a small number of SPD matrices per image set. These descriptors suffer from some drawbacks. One such drawback is the curse of dimensionality as each SPD matrix descriptor has the dimensions $WH \times WH$ (or more) assuming the images have size $W \times H$ (the descriptor is 160,000-D for images as small as 20×20). The image set may also contain too few images to reliably compute such high-dimensional descriptors, leading to undersampling at the level of each descriptor. Undersampling at the gallery level (and subsequent overfitting) may also be a problem as the typical gallery contains few image sets per class. This results in a correspondingly few number of high-dimensional descriptors that may not be enough to reliably train a machine learning model.

To avoid these problems, we extract a symmetric matrix feature $\mathbf{L} \in \mathbf{T}_{\mathbf{I}}\mathbb{S}_+^Q$ from each image $\boldsymbol{I}$, and we use a compressed version of its vectorization $\boldsymbol{a} = \operatorname{comp}(\operatorname{vec}(\mathbf{L})) \in \mathbb{R}^D$ as an atom in a dictionary corresponding to the image set. The steps for computing the image-level features are summarized in Fig. 2 and described below.

At the heart of our image-level descriptor is the use of Region Covariance Matrices (RCMs) [Pang *et al.*, 2008]. This is justified by the ability of covariances to fuse various types of features and keep track of their statistics. In addition, the covariance of a set of samples is invariant to rearrangement of these samples, giving the covariance more robustness to misalignment, a problem that face identification systems have to deal with even when automatic face alignment is applied, as the detection and alignment algorithms are still not perfect.

To compute the covariance matrices, we first compute a feature image $\boldsymbol{\phi}(\boldsymbol{I})$ similar to [Pang *et al.*, 2008; Harandi *et al.*, 2012] which produces at each pixel the following $M = 43$ responses:

$$\boldsymbol{\phi}_{x,y}^T = [\ x, y, \boldsymbol{I}(x,y), |G_{0,0}(x,y)|, \quad \ldots, \quad |G_{4,7}(x,y)|\ ]$$

where $G_{u,v}(x,y) = g_{u,v}(x,y) * \boldsymbol{I}(x,y)$ is the response of the image to the 2D Gabor wavelet $g_{u,v}(x,y)$ [Harandi *et al.*, 2012]:

$$\frac{k_v^2}{4\pi^2} e^{-\frac{k_v^2}{8\pi^2}(x^2+y^2)} \left(e^{ik_v(x\cos\theta_u + y\sin\theta_u)} - e^{-2\pi^2} \right)$$

where u is the orientation index, v is the scale index, $k_v = 1/\sqrt{2^{v-1}}$, and $\theta_u = \pi u/8$. To balance the trade-off between robustness to misalignment and spatial encoding, we follow the tradition of breaking the image into $n_r \times n_c$ cells and compute a covariance matrix for each cell based on the pixel responses in that cell.

To avoid the curse of dimensionality in the extracted descriptor, we compress the M responses at each pixel in cell (r,c), prior to computing the cell-specific covariance matrix, by projecting the M-D response vector into a subspace of a lower dimensionality m using a cell-specific, $M \times m$ column-orthogonal projection matrix $\mathbf{U}_{r,c}$. Each matrix $\mathbf{U}_{r,c}$ is computed by performing PCA on the M-D response vectors at all

pixels within the cell (r, c) from all the images in all gallery image sets. In our experiments, we set $m = 10$.

After compressing the responses, we calculate the $m \times m$ covariance matrix $\mathbf{\Sigma}_{r,c}$ from the m-D responses in cell (r, c). Next, we arrange the $n_r \times n_c$ covariances into the diagonal blocks of a $Q \times Q$ matrix $\mathbf{\Sigma}$, where $Q = n_r n_c m$. The matrix $\mathbf{\Sigma}$ can be easily shown to be SPD and so it lives in the non-Euclidean SPD manifold $\mathbb{S}_+^Q$. To measure the similarity in this non-Euclidean space, we endow $\mathbb{S}_+^Q$ with the LE Metric [Arsigny *et al.*, 2007] which measures the distance between any pair of SPD matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ by first using the Log map: $\text{Log} : \mathbb{S}_+^Q \rightarrow \mathbf{T_I}\mathbb{S}_+^Q$ to map them to the LE tangent space $\mathbf{T_I}\mathbb{S}_+^Q$ and then computing the Frobenius distance $\|\text{Log}\,\mathbf{X}_2 - \text{Log}\,\mathbf{X}_1\|_F$. If the Singular Value Decomposition (SVD) of an SPD matrix of dimensions $m \times m$ is $\mathbf{X} = \mathbf{U}\,\text{diag}(s_1, ..., s_m)\mathbf{V}^T$, the Log map is defined as:

$$\text{Log}\,\mathbf{X} = \mathbf{U}\,\text{diag}(\log s_1, ..., \log s_m)\mathbf{V}^T \tag{1}$$

The LE tangent space $\mathbf{T_I}\mathbb{S}_+^Q$ is equivalent to the space of symmetric matrices $\mathbb{S}^Q$, which is a vector space. This allows us to apply the familiar Euclidean machine learning algorithms to SPD matrices once they are mapped to the LE tangent space. Accordingly, the final steps are (a) mapping the SPD matrix $\mathbf{\Sigma}$ to the LE tangent space by computing $\mathbf{L} = \text{Log}\,\mathbf{\Sigma}$, (b) computing the uncompressed atom $\tilde{\boldsymbol{a}} = \text{vec}\,(\mathbf{L}) \in \mathbb{R}^{Q^2}$, and then obtaining the compressed atom $\boldsymbol{a} = \text{comp}\,(\tilde{\boldsymbol{a}}) \in \mathbb{R}^D$ where the operator comp() retains only the $D = n_r n_c m^2$ entries of $\tilde{\boldsymbol{a}}$ corresponding to the $n_r n_c$ diagonal blocks of $\mathbf{L}$ while discarding the rest (see the structure of $\boldsymbol{a}$ in Fig. 2).

Arranging the cell covariances into the diagonal blocks of $\mathbf{\Sigma}$ and mapping $\mathbf{\Sigma}$ to the LE tangent space unnecessarily requires more memory and processing time. Instead, we apply the equivalent but more efficient process of separately mapping each cell covariance matrix $\mathbf{\Sigma}_{r,c}$ to the LE tangent space, which gives $\mathbf{L}_{r,c} = \text{Log}\,\mathbf{\Sigma}_{r,c}$. In addition, we store only the $D = n_r n_c m^2$ nonzero entries of $\mathbf{L}$, which correspond to its diagonal blocks $\mathbf{L}_{r,c}$, into the compressed atom $\boldsymbol{a} \in \mathbb{R}^D$. All the remaining steps use the compressed atom $\boldsymbol{a}$ instead of the uncompressed, higher dimensional atom $\tilde{\boldsymbol{a}} \in \mathbb{R}^{Q^2}$.

The feature extraction step can be very efficiently implemented by making use of GPUs for performing convolutions and matrix multiplication. For each image, $n_r \times n_c$ small eigenvalue problems need to be computed for SPD matrices of size $m \times m$ in order to compute their matrix logarithms. Additional $n_r \times n_c$ eigenvalue problems of $M \times M$ matrices need to be solved for performing PCA during training but these are done only once for the complete gallery set rather than for each image.

3.2 Log-Euclidean Feature Learning

The goal of this step is to map the image descriptors from the LE tangent space $\mathbf{T_I}\mathbb{S}_+^Q$ into a lower-dimensional LE tangent space $\mathbf{T_I}\mathbb{S}_+^q$ in which they have a more discriminative subspace structure. In other words, we want the samples from one class to stay, in the new space, as far as possible from other-class subspaces while staying close to the same-class subspaces. In this new space, the sparse coding of a query sample $\boldsymbol{y}$ from class c over the dictionary $\mathbf{A}$ will more likely find that the subdictionary $\mathbf{A}_c$ provides better reconstruction of $\boldsymbol{y}$ compared to other subdictionaries. Consequently, the sparse coding will more likely associate $\boldsymbol{y}$ with its true class c.

Tangent Map Formulation: There are different ways to formulate the tangent map $\mathcal{W} : \mathbf{T_I}\mathbb{S}_+^Q \rightarrow \mathbf{T_I}\mathbb{S}_+^q$. One way is by the linear formulation given by:

$$\text{vec}\,(\mathbf{L}') = \mathcal{W}_1(\mathbf{L}) = \mathbf{W}^T\,\text{vec}\,(\mathbf{L}) \tag{2}$$

where $\mathbf{W} \in \mathbb{R}^{Q^2 \times q^2}$. To guarantee that $\mathbf{L}' \in \mathbb{S}^q$ for any $\mathbf{L} \in \mathbb{S}^Q$, the matrix $\mathbf{W}$ has to be constrained, such that it has $q(q+1)/2$ unique columns while the other $q(q-1)/2$ columns are permutations of other columns[1].

The second formulation $\mathcal{W}_2$ is a variation of $\mathcal{W}_1$ that avoids placing constraints on $\mathbf{W}$ by keeping only the unique $q(q+1)/2$ columns in $\mathbf{W}$ so that we just compute the (vectorized) lower triangular submatrix $\text{tril}(\mathbf{L}') \in \mathbb{R}^{q(q+1)/2}$ instead of the complete matrix $\mathbf{L}' \in \mathbb{R}^{q \times q}$:

$$\text{tril}(\mathbf{L}') = \mathcal{W}_2(\mathbf{L}) = \mathbf{W}^T\,\text{vec}\,(\mathbf{L}) \tag{3}$$

where $\mathbf{W} \in \mathbb{R}^{Q^2 \times q(q+1)/2}$. Since $\tilde{\boldsymbol{a}} = \text{vec}\,(\mathbf{L})$ has only D nonzero entries at known locations, the projection matrix $\mathbf{W}$ in both $\mathcal{W}_1$ and $\mathcal{W}_2$ needs only to contain the D rows corresponding to these nonzero entries. In this case, the dimensions of $\mathbf{W}$ in $\mathcal{W}_1$ can be reduced to $D \times q^2$ while in $\mathcal{W}_2$ it will be $D \times q(q+1)/2$. For simplicity, we use the second formulation, in which $\mathbf{W}$ is unconstrained and has dimensions $D \times q(q+1)/2$.

It is worth noting that a third formulation was used in [Xu *et al.*, 2015; Huang *et al.*, 2015b] which has the advantage of using much fewer parameters in the projection $\mathbf{W}$. However, the formulation is quadratic in the projection parameters compared to linear formulations $\mathcal{W}_1$ and $\mathcal{W}_2$. The quadratic formulation is useful for applications in which the SPD descriptors are very high-dimensional such as the 400×400 image set covariance used by Wang *et al.* [2012]. The SPD descriptors in this paper have considerably fewer dimensions, and so we opt to use the simpler linear form $\mathcal{W}_2$. As we see later, our choice of a formulation that is linear in the parameters leads to an easier-to-solve optimization problem in which finding the globally optimal solution is straightforward and efficient.

Optimization Problem: Let $\mathbf{A}_c \in \mathbb{R}^{D \times N_c}$ be the dictionary containing all the N_c compressed atoms from all image sets associated with class c (after removing all identical atoms due to identical images):

$$\mathbf{A}_c = [\ \boldsymbol{a}_1 \quad \boldsymbol{a}_2 \quad \ldots \quad \boldsymbol{a}_{N_c}\ ]$$

Furthermore, let N be the total number of atoms in the gallery, C be the number of classes, $c(i)$ be the class associated with atom $\boldsymbol{a}_i$, and let $\boldsymbol{z}_{i,c}$ be the dense representation

[1]There are other algebraically equivalent ways to express the constraint on the columns of $\mathbf{W}$, all of them leading to the same measure of distance between symmetric matrices. Since we do not use the formulation $\mathcal{W}_1$ in this paper, further elaboration on these ways is beyond the scope of this paper.

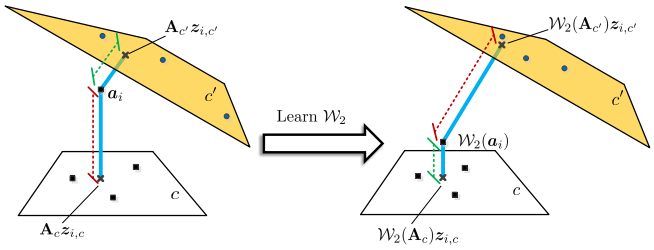

Figure 3: To improve the discriminative subspace arrangement of the data, the LE feature map $\mathcal{W}_2$ is learned such that it maximizes the distance between each atom $\boldsymbol{a}_i$ and its projection $\mathbf{A}_{c'}\boldsymbol{z}_{i,c}$ on every other-class dictionary $\mathbf{A}_{c'}$ while minimizing the distance between the sample and its projection $\mathbf{A}_c\boldsymbol{z}_{i,c}$ on the dictionary $\mathbf{A}_c$ of its own class c.

of an atom $\boldsymbol{a}_i$ with respect to the dictionary $\mathbf{A}_c$ of a different class c:

$$\boldsymbol{z}_{i,c} = \underset{\boldsymbol{z}}{\operatorname{argmin}} \|\boldsymbol{a}_i - \mathbf{A}_c\boldsymbol{z}\|_2^2 + \lambda_1 \|\boldsymbol{z}\|_2^2 \tag{4}$$

where we use $\lambda_1 = 0.001$. If we let $\mathbf{J}_c = \mathbf{A}_c^T\mathbf{A}_c + \lambda_1\mathbf{I}$, we obtain $\boldsymbol{z}_{i,c} = \mathbf{J}_c^{-1}\mathbf{A}_c^T\boldsymbol{a}_i$. The first goal we need the tangent map $\mathcal{W}_2$ to achieve is to maximize the distance between every atom $\boldsymbol{a}_i$, from a certain class $c(i)$, and its dense projection $\mathbf{A}_c\boldsymbol{z}_{i,c}$ on the dictionary of each other class $c \neq c(i)$ (see Fig. 3):

$$\frac{1}{C}\sum_{c=1}^{C}\sum_{i,c(i)\neq c}\frac{1}{N_{c(i)}(C-1)}\left\|\mathbf{W}^T(\boldsymbol{a}_i - \mathbf{A}_c\boldsymbol{z}_{i,c})\right\|_2^2 \tag{5}$$

$$= \operatorname{tr}\mathbf{W}^T\mathbf{S}_1\mathbf{W} \tag{6}$$

where $\mathbf{S}_1 = \frac{1}{C}\sum_{c=1}^{C}\sum_{i,c(i)\neq c}\frac{1}{N_{c(i)}(C-1)}(\boldsymbol{a}_i - \mathbf{A}_c\boldsymbol{z}_{i,c})(\boldsymbol{a}_i - \mathbf{A}_c\boldsymbol{z}_{i,c})^T$.

The reason we use dense, l_2-regularized representations is that it has a closed form solution that is more efficient to evaluate. Moreover (and more importantly), we want to maximize the distance between each atom $\boldsymbol{a}_i$ and the span of as many as possible of those different-class atoms that may contribute to reconstructing $\boldsymbol{a}_i$. This makes the dense representation a more appropriate choice.

Before describing the other goal, we redefine the dense representation $\boldsymbol{z}_{i,c}$ for the case in which we project $\boldsymbol{a}_i$ on the dictionary of its own class (i.e. $c = c(i)$):

$$\boldsymbol{z}_{i,c} = \underset{\boldsymbol{z}}{\operatorname{argmin}} \|\boldsymbol{a}_i - \mathbf{A}_c\boldsymbol{z}\|_2^2 + \lambda_1 \|\boldsymbol{z}\|_2^2, \text{s.t. } z^{(i)} = 0. \tag{7}$$

The only difference between (4) and (7) is the constraint $z^{(i)} = 0$ which excludes any solution in which $\boldsymbol{a}_i$ contributes to its own representation. If we let $\boldsymbol{u}_{i,c} = \mathbf{J}_c^{-1}\mathbf{A}_c^T\boldsymbol{a}_i$, and $w_i = \boldsymbol{u}_{i,c}^{(i)}/\mathbf{J}_c^{-1(i,i)}$, we obtain $\boldsymbol{z}_{i,c} = \boldsymbol{u}_{i,c} - w_i\text{col}_i(\mathbf{J}_c^{-1})$.

The other goal the tangent map has to meet is minimizing the distance between every atom $\boldsymbol{a}_i$ from a certain class c to its dense projection $\mathbf{A}_c\boldsymbol{z}_{i,c}$ on the dictionary of its own class:

$$\frac{1}{C}\sum_{c=1}^{C}\frac{1}{N_c}\sum_{i,c(i)=c}\left\|\mathbf{W}^T(\boldsymbol{a}_i - \mathbf{A}_c\boldsymbol{z}_{i,c})\right\|_2^2 \tag{8}$$

$$= \operatorname{tr}\mathbf{W}^T\mathbf{S}_2\mathbf{W} \tag{9}$$

where $\mathbf{S}_2 = \sum_{c=1}^{C}\frac{1}{N_c}\sum_{i,c(i)=c}(\boldsymbol{a}_i - \mathbf{A}_c\boldsymbol{z}_{i,c})(\boldsymbol{a}_i - \mathbf{A}_c\boldsymbol{z}_{i,c})^T$. In addition, we add a regularization term $\|\mathbf{W}\|_F^2 = \operatorname{tr}\mathbf{W}^T\mathbf{W}$ to the quantity to be minimized. We then combine all goals in one criterion by maximizing the following ratio of quadratic forms:

$$\max_{\mathbf{W}}\frac{\operatorname{tr}\mathbf{W}^T\mathbf{S}_1\mathbf{W}}{\operatorname{tr}\mathbf{W}^T(\mathbf{S}_2+\mathbf{I})\mathbf{W}} \tag{10}$$

The optimal solution to this problem is obtained by finding the $q(q+1)/2$ generalized eigenvectors with the largest eigenvalues of the following generalized eigenvalue problem:

$$\mathbf{S}_1\boldsymbol{w}_k = \lambda_k(\mathbf{S}_2+\mathbf{I})\boldsymbol{w}_k$$

After finding $\mathbf{W}$, we use it to embed the dictionaries of all classes. If we assume all the C classes have the same number of images $N_c = N/C$, the computational complexity of feature learning is $O\left(D^3 + C\times\left(CN_c^3 + DN_c^2 + CD^2N_c\right)\right)$, where it takes $O(D^3)$ for the solution of the $D\times D$ generalized eigenvalue problem in (10), $O(DN_c^2+N_c^3)$ for computing $\mathbf{J}_c$ and inverting it for one class, $O(CN_c^3)$ for computing the representations of same-class samples and other-class samples with respect to the dictionary $\mathbf{A}_c$ of one class, and $O(CD^2N_c)$ for computing the contribution of one class to the two scatter matrices $\mathbf{S}_1$ and $\mathbf{S}_2$.

3.3 Coding and Classification

Given a probe image set $\mathbf{Y} = [\boldsymbol{y}_1,\ldots,\boldsymbol{y}_{N_y}]$, the method extracts the LE dictionary from the set $\mathbf{Y}$ as described in Section 3.1; then uses the tangent map $\mathbf{W}$ to project each atom in $\mathbf{Y}$'s dictionary to the LE feature space. Subsequently, we apply SRC to compute the label for $\mathbf{Y}$. More specifically, we solve for the sparse representation vector $\boldsymbol{x}\in\mathbb{R}^N$ corresponding to the mean $\bar{\boldsymbol{y}}$ of the embedded feature vectors (i.e. LE Frechet mean [Arsigny *et al.*, 2007]):

$$\boldsymbol{x} = \underset{\boldsymbol{x}\in\mathbb{R}^N}{\operatorname{argmin}}\frac{1}{2}\|\bar{\boldsymbol{y}} - \mathbf{A}\boldsymbol{x}\|^2 + \lambda\|\boldsymbol{x}\|_1$$

where $\mathbf{A}$ is the dictionary containing all the embedded LE atoms from all classes. Given the sparse representation $\bar{\boldsymbol{x}}$, we can find the class contributing the most to the representation, and with which $\bar{\boldsymbol{y}}$ should be associated, using the minimum residual rule of [Wright *et al.*, 2009]. If we let δ_c be the $N\times N$ diagonal matrix with all zeros except at the N_c diagonal entries corresponding to the atoms of class c, the residual $r_c(\boldsymbol{y};\bar{\boldsymbol{x}})$ corresponding to class c is given by:

$$r_c(\boldsymbol{y};\bar{\boldsymbol{x}}) = \|\boldsymbol{y} - \mathbf{A}\delta_c\bar{\boldsymbol{x}}\|^2 \tag{11}$$

The class for which r_c is minimum is chosen as the label for the probe set.

4 Experimental Evaluation

We have conducted extensive experiments to compare the performance of the proposed algorithm, i.e. Log-Euclidean Feature Learning with SRC (LEFL-SRC), against several existing algorithms for image-set classification. The compared methods include Affine Hull-based Image Set Distance

Figure 4: Sample frames from YTC, YTF and MobFaces. Each column shows two photos of the same person from the same dataset. The photos in the first, second and third columns are from YTC, YTF, and MobFaces, respectively. YTC and YTF photos reveal the large intra-class appearance variations present in both datasets. MobFaces photos are relatively frontal but they reveal various challenges such as occlusion and blur in addition to other significant intra-class variations in illumination and context due to the change in sessions. The bottom-right MobFaces photo is best viewed on screen.

(AHISD) [Cevikalp and Triggs, 2010], its convex variant (CHISD) [Cevikalp and Triggs, 2010], Sparse-Approximated Nearest Points (SANP) [Hu *et al.*, 2011], Dictionary-based Face Recognition from Videos (DFRV) [Chen *et al.*, 2012], Mean Sequence Sparse Representation-based Classification (MS-SRC) [Ortiz *et al.*, 2013], a variation of MS-SRC that uses Collaborative Representation-based Classification (CRC) [Zhang *et al.*, 2011] for classifying the mean of the sequence (MS-CRC), Set to Set Distance Metric Learning (SSDML) [Zhu *et al.*, 2013], Deep Reconstruction Models (DRM) [Hayat *et al.*, 2014a], Projection Metric Learning (PML) [Huang *et al.*, 2015a], and Log-Euclidean Metric Learning (LEML) [Huang *et al.*, 2015b]. To understand the contribution to performance made by the different components of our classifier, we are also comparing with two variants of our classifier: one without LE features but with feature learning applied to intensity features (FL-SRC), and another with the LE features but without the feature learning (LE-SRC).

For existing methods, we have used the source code provided by the original authors and set the parameters according to the recommendations made in their respective papers. The only exception to this are MS-CRC and MS-SRC which we have implemented ourselves. To guarantee a fair comparison, the same features and dataset splits were used to compare all the methods. We made an exception for the DRM approach where we report the performance using the 1475-D LBP features extracted from the intensity features used with the rest of the methods. The reason for this exception is that the original paper of DRM [Hayat *et al.*, 2014a] and its publicly available source-code included the extraction of LBP features as one of the preprocessing steps of DRM. For PML, we have modified the method to deal with the situation in which the number of images n_s in a given image set s is lower than the dimensionality d of the subspace PML computes from each image set. In that case, we synthesize additional images by small random translations and rotations of the original n_s images so that s has $2d$ images. Since PML requires the gallery to have at least two image sets for each class whereas MobFaces-I provides a single gallery image set per class, we randomly split each gallery set in MobFaces-I into two subsets of nearly equal sizes (the difference in size is at most one).

Parameter Settings: We use the following parameters in our proposed method. We resize each input image to $w = 120$ and $h = 144$. We then divide each image into a $n_r = 6 \times n_c = 6$ grid of non-overlapping cells for calculating the RCMs. As stated earlier, we use $m = 10$. Finally, we set the dimension of the lower-dimensional LE tangent space to $q = 28$ which corresponds to an LE tangent map $\mathbf{W}$ with $q(q+1)/2 = 406$ columns. It is worth noting that smaller grids (i.e. smaller n_r and n_c) lead to inferior recognition performance. Grids larger than $n_r = 6 \times n_c = 6$ could possibly lead to better performance at the expense of increasing the memory footprint of the algorithm. However, we have not tried such larger grids.

4.1 YouTube Celebrities (YTC)

The YTC dataset contains 1,910 YouTube-downloaded videos of 47 subjects [Kim *et al.*, 2008]. For a given subject, the videos are short segments clipped from three longer, parent videos downloaded from YouTube. YTC has been built to be very challenging for face tracking and recognition by choosing low resolution videos with wild variations in pose, scale, hair style, make-up, illumination, motion and number of people per frame.

Experimental Protocol: We run ten-fold cross-validation experiment. The $9 \times 47 = 423$ videos in each fold are randomly selected from the complete dataset while minimizing the overlap between different folds as much as possible.

Feature Extraction: We use the Viola-Jones (VJ) detector [Viola and Jones, 2004] to locate the faces in each video. Then we use the eye locations detected using the method of Asthana *et al.* [2013] to align the subject's face to a standard, 30×36 pixel frame. The intensities are histogram equalized and arranged in a 1080-D feature vector. We use the feature vectors from a given video define the corresponding image set.[2]

4.2 YouTube Faces (YTF)

The YTF dataset contains 3,425 videos of 1,595 subjects with diverse ethnicities [Wolf *et al.*, 2011]. Similar to YTC, YTF videos are downloaded from YouTube and are very challenging for face recognition. We conduct our experiments on those subjects with four or more videos available. This results in 226 subjects. After randomly dropping one subject, we randomly split the remaining 225 subjects into five mutually exclusive groups, with 45 subjects each. We repeat the

[2]We have not cleaned any of the bad detections or misaligned faces in an effort to test the robustness of the compared methods to such outliers.

Table 1: The multi-fold sample mean and standard deviation of the recognition rates obtained with the compared methods on YTC and YTF. We have highlighted in bold the rates of the top two performing methods for each dataset. Although YTC and YTF have similar challenges, the rates obtained for YTC are higher because the test protocol for YTC guarantees that for each test video clip there is a corresponding gallery video clip such that both are segments from the same parent YouTube video.

Methods	YTC	YTF
AHISD	57.27 ± 3.44	17.18 ± 8.93
CHISD	64.79 ± 1.72	32.99 ± 7.97
SANP	66.99 ± 0.69	31.62 ± 8.56
DFRV	66.70 ± 1.52	36.77 ± 10.19
MS-CRC	66.88 ± 2.21	43.64 ± 8.27
MS-SRC	74.68 ± 1.96	45.02 ± 5.82
SSDML	69.22 ± 1.64	34.02 ± 10.03
DRM	70.35 ± 2.52	43.99 ± 5.23
PML	68.55 ± 1.76	40.21 ± 11.98
LEML	60.32 ± 1.80	30.93 ± 2.55
LE-LEML	73.26 ± 1.50	48.45 ± 5.66
FL-SRC (ours)	$\mathbf{75.71 \pm 1.57}$	45.36 ± 3.45
LE-SRC (ours)	75.11 ± 1.49	$\mathbf{49.83 \pm 7.51}$
LEFL-SRC (ours)	$\mathbf{76.28 \pm 2.22}$	$\mathbf{53.26 \pm 8.10}$

experiment for each group where we use the first three videos of each subject as gallery sets and the remaining videos for testing. We stress that we run face identification (multi-classification) experiments rather than binary verification as is usually done on YTF in the literature. Since the dataset provides aligned face images, we extract intensity features from each image by cropping the central 100×100 box from each image, resizing it to 30×36, and histogram-equalizing it.

4.3 Mobile Faces (MobFaces)

The MobFaces dataset contains 750 videos of 50 subjects taken by a smartphone's front camera during various user interactions with the phone [Fathy *et al.*, 2015]. There are three sessions of five videos each (one enrollment + four tasks) per subject where each session is taken under a different illumination and/or in a different place. The dataset includes some of the unique challenges of mobile-based continuous facial authentication such as wild variations in illumination and context due to the mobility of the device. We compute the features using the same pipeline we developed for the YTC dataset. Although the features used in this paper are different from [Fathy *et al.*, 2015], we adopt the two evaluation protocols suggested in [Fathy *et al.*, 2015] by dividing the task videos into ten-second long clips and treating each clip as a separate query. In the first protocol (MobFaces-I), training is done using only the 50 enrollment videos of one session and testing is performed on the ten-second long task video clips from the two other sessions. In the second protocol (MobFaces-II), training is done on the 100 enrollment videos of two sessions and testing is done on the task video clips of the remaining session. Results are reported for each of the six scenarios possible with these protocols. The clipping of the 600 task videos results in 1065 ten-second clips for the first session, 587 the second, and 666 for the third.

4.4 Results

Table 1 shows the mean and standard-deviation of the recognition rates of the compared methods for the YTC and YTF datasets while Table 2 shows the recognition rates for the six different evaluation scenarios for the MobFaces dataset. Both tables clearly show the superiority of the proposed method (LEFL-SRC) in comparison with other methods.

The improvement in performance on MS-SRC by FL-SRC is not significant except on YTC. This is because feature learning does not help much with intensity features, which inherently have a good discriminative subspace structure. Accordingly, the only significant advantage FL-SRC provides over MS-SRC is the reduction of dimensionality without loss in identification accuracy. On the other hand, the results show that the improvement due to feature learning is significant when we compare LE-SRC with LEFL-SRC. Since LE features are nonlinear in intensities, they do not preserve the sparse linear dependencies between samples though these nonlinear features are more robust. In addition to reducing the dimensionality of the LE features, the feature learning algorithm improves their discriminative subspace structure which in turn boosts the performance of SRC. It is worth noting that although LEFL-SRC outperforms MS-SRC and LE-SRC, LEFL-SRC uses only 406 features per atom which is fewer than the $30 \times 36 = 1080$ features used by MS-SRC and the $D = 3600$ features used by LE-SRC.

We emphasize that we are using a slightly different experimental pipeline. In particular, we detect face landmarks and use them to align the detected faces. While this should generally improve the performance of all methods, this is not the case for LEML [Huang *et al.*, 2015b] as alignment makes many of the face images in each set near identical, resulting in a near-zero image set covariance, which is the major component of LEML's Gaussian descriptor. Such degeneracy explains the relatively low accuracy obtained for LEML. We have also derived a modified version of LEML that we refer to as LE-LEML. It uses the image-level SPD descriptors proposed in this paper instead of the set-level Gaussian descriptors. Tables 1 and 2 clearly show the use of the image-level descriptors significantly boosts the performance of LEML.

5 Conclusion

We proposed LEFL-SRC, an approach for face identification from image sets using sparse coding on the Log-Euclidean (LE) tangent space of the SPD manifold $\mathbb{S}_+^Q$. In this approach, we first describe each image by generating a number of local covariance descriptors arranged in a grid; then we use a dictionary of atoms from the LE tangent space $\mathrm{T}_{\mathbf{I}}\mathbb{S}_+^Q$ to represent each gallery image set. While previous LE-based approaches for image set classification extract a single or very few LE samples with very high dimensionality, our approach extracts from each set many LE samples of much lower dimensionality. We formulated an optimization problem for learning an embedding into a lower-dimensional LE tangent space $\mathrm{T}_{\mathbf{I}}\mathbb{S}_+^q$

Table 2: The recognition rates obtained on the MobFaces dataset under the different protocols. The setting $(1 \rightarrow \{2,3\})$ involves training on session 1 (i.e. the lit session) and testing on sessions 2 and 3 (i.e. the unlit and day-lit sessions). The other five settings are defined in a similar manner. Each 'avg' column contains the average of the rates obtained under the three settings to its left. Since each session has a different number of test video clips, the average column weighs the rate of each setting by its number of test sets. We have highlighted in bold the rates of the top two performing methods for each setting. Note that the row for DRM-GRAY corresponds to the performance of the DRM method obtained when we use intensity features instead of the LBP features DRM calculates by default during feature preprocessing. While intensity features lead to noticeable performance improvement for DRM on MobFaces, they decrease the performance of DRM on the YTC dataset compared to LBP features and so we report only for YTC the performance of DRM on the LBP features.

	MobFaces-I				MobFaces-II			
Methods	$1 \rightarrow \{2,3\}$	$2 \rightarrow \{1,3\}$	$3 \rightarrow \{1,2\}$	avg	$\{2,3\} \rightarrow 1$	$\{1,3\} \rightarrow 2$	$\{1,2\} \rightarrow 3$	avg
AHISD	15.00	31.14	29.30	26.12	24.41	51.28	52.85	39.39
CHISD	10.61	26.57	25.73	21.96	23.29	44.97	47.60	35.76
SANP	9.34	27.09	26.15	21.96	20.38	48.89	45.95	34.94
DFRV	19.39	32.29	30.87	28.30	32.11	50.60	52.40	42.62
MS-CRC	**48.20**	51.30	50.24	50.09	**69.01**	**73.59**	77.18	**72.52**
MS-SRC	32.40	46.56	42.49	41.29	43.29	71.89	75.53	59.79
SSDML	10.53	28.89	26.15	22.95	21.31	50.09	54.95	38.27
DRM-LBP	23.46	32.41	36.38	31.41	38.97	62.86	65.77	52.72
DRM-GRAY	33.28	38.94	37.95	37.06	53.62	70.53	69.37	62.42
PML	**51.16**	45.98	41.77	45.88	45.92	56.56	61.41	53.06
LEML	13.17	20.80	21.19	18.87	17.37	29.47	33.03	24.94
LE-LEML	42.70	45.93	44.07	44.39	49.39	66.95	74.62	61.09
FL-SRC (ours)	32.88	46.97	42.25	41.48	44.98	72.40	76.58	61.00
LE-SRC (ours)	47.01	**52.63**	**54.00**	**51.60**	59.25	**73.59**	**84.98**	70.28
LEFL-SRC (ours)	48.20	**56.21**	**54.90**	**53.58**	**62.72**	**75.64**	**86.19**	**72.74**

in which the data has a more discriminative subspace structure allowing accurate subsequent application of SRC. Extensive experiments on three public datasets (YTC, YTF, and MobFaces) show that our method outperforms many state-of-the-art methods. In addition, we made an empirical ablation analysis where we have shown how the different components of our approach contribute to the final perofrmance.

Acknowledgments

This work was supported by a cooperative agreement FA8750-13-2-0279 from DARPA. We thank Pouya Samangouei for helpful discussions and the anonymous reviewers for their valuable feedback.

References

[Arsigny *et al.*, 2007] Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM j. on matrix analysis and applications*, 29(1):328–347, 2007.

[Asthana *et al.*, 2013] Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, and Maja Pantic. Robust discriminative response map fitting with constrained local models. In *CVPR*, pages 3444–3451, 2013.

[Basri and Jacobs, 2003] Ronen Basri and David W Jacobs. Lambertian reflectance and linear subspaces. *PAMI*, 25(2):218–233, 2003.

[Cevikalp and Triggs, 2010] Hakan Cevikalp and Bill Triggs. Face recognition based on image sets. In *CVPR*, pages 2567–2573, 2010.

[Chen *et al.*, 2012] Yi-Chen Chen, Vishal M Patel, P Jonathon Phillips, and Rama Chellappa. Dictionary-based face recognition from video. In *ECCV*, pages 766–779. 2012.

[Chen *et al.*, 2013] Shaokang Chen, Conrad Sanderson, Mehrtash T Harandi, and Brian C Lovell. Improved image set classification via joint sparse approximated nearest subspaces. In *CVPR*, pages 452–459, 2013.

[Fathy *et al.*, 2015] Mohammed E Fathy, Vishal M Patel, and Rama Chellappa. Face-based active authentication on mobile devices. In *ICASSP*, pages 1687–1691, 2015.

[Guo *et al.*, 2010] Kai Guo, Prakash Ishwar, and Janusz Konrad. Action recognition using sparse representation on covariance manifolds of optical flow. In *AVSS*, pages 188–195, 2010.

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

[Harandi *et al.*, 2011] Mehrtash T Harandi, Conrad Sanderson, Sareh Shirazi, and Brian C Lovell. Graph embedding discriminant analysis on grassmannian manifolds for improved image set matching. In *CVPR*, pages 2705–2712, 2011.

[Harandi *et al.*, 2012] Mehrtash T Harandi, Conrad Sanderson, Richard Hartley, and Brian C Lovell. Sparse coding

and dictionary learning for symmetric positive definite matrices: A kernel approach. In *ECCV*, pages 216–229, 2012.

[Harandi *et al.*, 2013] Mehrtash Harandi, Conrad Sanderson, Chunhua Shen, and Brian C Lovell. Dictionary learning and sparse coding on grassmann manifolds: An extrinsic solution. In *ICCV*, pages 3120–3127, 2013.

[Hayat *et al.*, 2014a] Munawar Hayat, Mohammed Bennamoun, and Senjian An. Learning non-linear reconstruction models for image set classification. In *CVPR*, pages 1915–1922, 2014.

[Hayat *et al.*, 2014b] Munawar Hayat, Mohammed Bennamoun, and Senjian An. Reverse training: An efficient approach for image set classification. In *ECCV*, pages 784–799. 2014.

[Hu *et al.*, 2011] Yiqun Hu, Ajmal S Mian, and Robyn Owens. Sparse approximated nearest points for image set classification. In *CVPR*, pages 121–128, 2011.

[Huang *et al.*, 2015a] Zhiwu Huang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Projection metric learning on grassmann manifold with application to video based face recognition. In *CVPR*, pages 140–149, 2015.

[Huang *et al.*, 2015b] Zhiwu Huang, Ruiping Wang, Shiguang Shan, Xianqiu Li, and Xilin Chen. Log-euclidean metric learning on symmetric positive definite manifold with application to image set classification. In *ICML*, pages 720–729, 2015.

[Kim *et al.*, 2008] Minyoung Kim, Sanjiv Kumar, Vladimir Pavlovic, and Henry Rowley. Face tracking and recognition with visual constraints in real-world videos. In *CVPR*, pages 1–8, 2008.

[Li *et al.*, 2013] Peihua Li, Qilong Wang, and Lei Zhang. A novel earth mover's distance methodology for image matching with gaussian mixture models. In *ICCV*, pages 1689–1696, 2013.

[Lu *et al.*, 2015] Jiwen Lu, Gang Wang, Weihong Deng, Pierre Moulin, and Jie Zhou. Multi-manifold deep metric learning for image set classification. In *CVPR*, pages 1137–1145, 2015.

[Mahmood *et al.*, 2014] Arif Mahmood, Ajmal Mian, and Robyn Owens. Semi-supervised spectral clustering for image set classification. In *CVPR*, pages 121–128, 2014.

[Ortiz *et al.*, 2013] Enrique G Ortiz, Alan Wright, and Mubarak Shah. Face recognition in movie trailers via mean sequence sparse representation-based classification. In *CVPR*, pages 3531–3538, 2013.

[Pang *et al.*, 2008] Yanwei Pang, Yuan Yuan, and Xuelong Li. Gabor-based region covariance matrices for face recognition. *IEEE Trans. on Circuits and Systems for Video Technology*, 18(7):989–993, 2008.

[Qiu and Sapiro, 2015] Qiang Qiu and Guillermo Sapiro. Learning transformations for clustering and classification. *JMLR*, 16(1):187–225, 2015.

[Vemulapalli and Jacobs, 2015] Raviteja Vemulapalli and David W Jacobs. Riemannian metric learning for symmetric positive definite matrices. *arXiv preprint arXiv:1501.02393*, 2015.

[Viola and Jones, 2004] Paul Viola and Michael J Jones. Robust real-time face detection. *IJCV*, 57(2):137–154, 2004.

[Wang and Chen, 2009] Ruiping Wang and Xilin Chen. Manifold discriminant analysis. In *CVPR*, pages 429–436, 2009.

[Wang *et al.*, 2008] Ruiping Wang, Shiguang Shan, Xilin Chen, and Wen Gao. Manifold-manifold distance with application to face recognition based on image set. In *CVPR*, pages 1–8, 2008.

[Wang *et al.*, 2012] Ruiping Wang, Huimin Guo, Larry S Davis, and Qionghai Dai. Covariance discriminative learning: A natural and efficient approach to image set classification. In *CVPR*, pages 2496–2503, 2012.

[Wang *et al.*, 2015] Wen Wang, Ruiping Wang, Zhiwu Huang, Shiguang Shan, and Xilin Chen. Discriminant analysis on riemannian manifold of gaussian distributions for face recognition with image sets. In *CVPR*, pages 2048–2057, 2015.

[Wolf *et al.*, 2011] Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR*, pages 529–534, 2011.

[Wright *et al.*, 2009] John Wright, Allen Y Yang, Arvind Ganesh, Shankar S Sastry, and Yi Ma. Robust face recognition via sparse representation. *PAMI*, 31(2):210–227, 2009.

[Xu *et al.*, 2015] Chunyan Xu, Canyi Lu, Junbin Gao, Wei Zheng, Tianjiang Wang, and Shuicheng Yan. Discriminative analysis for symmetric positive definite matrices on lie groups. *IEEE Trans. on Circuits and Systems for Video Technology*, 25(10):1576–1585, 2015.

[Yger and Sugiyama, 2015] Florian Yger and Masashi Sugiyama. Supervised logeuclidean metric learning for symmetric positive definite matrices. *arXiv preprint arXiv:1502.03505*, 2015.

[Yuan *et al.*, 2010] Chunfeng Yuan, Weiming Hu, Xi Li, Stephen Maybank, and Guan Luo. Human action recognition under log-euclidean riemannian metric. In *ACCV*, pages 343–353. 2010.

[Zhang *et al.*, 2011] Lei Zhang, Meng Yang, and Xiangchu Feng. Sparse representation or collaborative representation: Which helps face recognition? In *ICCV*, pages 471–478, 2011.

[Zhu *et al.*, 2013] Pengfei Zhu, Lei Zhang, Wangmeng Zuo, and Dejing Zhang. From point to set: Extend the learning of distance metrics. In *ICCV*, pages 2664–2671, 2013.