A Supervised Low-rank Method for Learning Invariant Subspaces

Farzad Siyahjani Ranya Almohsen Sinan Sabri Gianfranco Doretto
West Virginia University
Morgantown, WV 26508
{fsiyahja, ralmohse, sisabri, gidoretto}@mix.wvu.edu

Abstract

Sparse representation and low-rank matrix decomposition approaches have been successfully applied to several computer vision problems. They build a generative representation of the data, which often requires complex training as well as testing to be robust against data variations induced by nuisance factors. We introduce the invariant components, a discriminative representation invariant to nuisance factors, because it spans subspaces orthogonal to the space where nuisance factors are defined. This allows developing a framework based on geometry that ensures a uniform inter-class separation, and a very efficient and robust classification based on simple nearest neighbor. In addition, we show how the approach is equivalent to a local metric learning, where the local metrics (one for each class) are learned jointly, rather than independently, thus avoiding the risk of overfitting without the need for additional regularization. We evaluated the approach for face recognition with highly corrupted training and testing data, obtaining very promising results.

1. Introduction

Recent approaches based on sparse representation [44] and low-rank matrix decomposition [5] have demonstrated great potential for addressing the problem of human identification, based on matching face images. Sparse coding has led to impressive performance even for image classification [30, 12]. Similarly, low-rank methods, after being applied to domains such as segmentation and grouping [7], tracking [49], and 3D visual recovery [26], are now also being used for classification [50]. For face recognition the sparse representation-based classification (SRC) method [44] has shown robustness with respect to a high degree of noise and occlusions in the test images. At the same time, sparse coding dictionary learning was shown to be sensitive to training samples corrupted by structural nuisance factors, such as occlusions, disguise, pose, lighting variations, and so on. This has motivated the development of low-rank matrix decomposition approaches [5, 6, 48, 39], which have the ability to learn a representational dictionary even in presence of corrupted data. Those approaches build a generative representation of the data that focusses on capturing all the information descriptive of an entity. This leads to complex training and testing for building robustness against, and filtering out unwanted data variations due to nuisance factors.

In this work we introduce a low-rank modeling framework that gives up capturing all the descriptive information of an entity (referred to as the *sufficient component*), and focusses on learning a representation that is invariant to nuisance factors (referred to as the *invariant component*). The main advantage of this approach is a fast procedure for computing and comparing invariant components for recognition. Indeed, we will see that this can be achieved by a simple matrix multiplication. On the other hand, the main challenge of this approach is that different entities may originate the same invariant component, thus preventing their discrimination. We will show that the proposed framework not only learns different invariant representations for different entities, but such representations promote a uniform inter-class separation.

The approach couples simple geometry tools with recent advances in low-rank matrix recovery theory [5], and develops a supervised model for learning the proposed invariant representation, which spans an *invariant subspace*. Such subspace has to be orthogonal to the *variation subspace*, generated by data variation induced by nuisance factors on all the entities. We make the assumption that the variation subspace is low-rank. Although this is an approximation, we empirically verify that it leads to very promising results for face recognition when training and testing data are highly corrupted, which is typical in video surveillance applications.

While the framework is grounded on geometry, we will show how it relates to metric learning [36, 9, 34, 4], typically used for improving nearest neighbor (NN) classification based on the Euclidean distance. We will show that learning the invariant components is equivalent to learning the representatives of a set of entities (or classes), thus classification is based on identifying the nearest invariant component. Less intuitively, the same invariant components define a *global* metric, and also a *local* metric. This is important because local metric learning approaches [16, 1, 13, 42], improve upon global ones by taking into account the variability of the discriminative power of features across different neighborhoods. In particular, most of the approaches learn local metrics for different neighborhoods independently, and use regularization to avoid overfitting. Our framework learns the invariant components, and therefore the local metrics jointly. In addition, their interpretation as a global metric is shown to promote uniform inter-class separation.

The rest of the paper will build more connections and differentiations with the related literature by taking advantage of the introduced notation. In addition, Section 2 introduces the idea of *invariant subspace*. Section 3 highlights its advantages and challenges for classification. Section 4 describes a supervised model for training. Section 5 shows how classification is done, describes its properties, and defines the global and local metrics being learned. Finally, Section 6 validates the proposed approach.

2. Invariant Subspace Representation

We assume that a data point $x \in \mathbb{R}^m$, representing an entity (e.g., the vectorized version of the image pixels of a face), can be modeled by two additive components. The first one, $s \in \mathbb{R}^m$, represents all the information necessary to recognize the entity (e.g., everything that describes the specific identity of the individual depicted by the face image). From a statistical point of view, we can imagine s to be the equivalent of a sufficient statistic for recognition, and we refer to it as the *sufficient component*. The second component, $v \in \mathbb{R}^m$, represents how the data point differs from the sufficient component by the effect of nuisance factors, which are not descriptive of the entity. For instance, the image of a face might be modified by different lighting conditions, facial expressions, occlusions, etc. We call *variation subspace*, $\mathcal{V} \subset \mathbb{R}^m$, the space where the *variation component* v is defined. It is assumed that v spans $\mathcal{V}$ as changes in nuisance factors affect a data point, which is modeled as

$$x \doteq s + v \ . \tag{1}$$

If $P_{\mathcal{V}} : \mathbb{R}^m \to \mathcal{V}$ is the projection operator mapping an m-dimensional vector onto $\mathcal{V}$, x can be further decomposed as $x = (P_{\mathcal{V}} s + v) + (s - P_{\mathcal{V}} s)$. In particular, the first component $a \doteq P_{\mathcal{V}} s + v$, is defined in $\mathcal{V}$, whereas the second component $b \doteq s - P_{\mathcal{V}} s$, is defined in the orthogonal complement of the variation space, $\mathcal{V}^{\perp}$.

The decomposition $x = a + b$ has the following property. Let us assume that x_1 and x_2 are two different points representing the same entity. According to (1), it must be that $x_1 = s + v_1$ and $x_2 = s + v_2$, because they have been affected by different nuisance factors. This means that $a_1 = P_{\mathcal{V}} s + v_1$, and $a_2 = P_{\mathcal{V}} s + v_2$; however, $b_1 = s - P_{\mathcal{V}} s = b_2$, which highlights that the component b is *invariant* to the changes induced by the nuisance factors. We refer to the subspace where b is defined as the *invariant subspace*[1] $\mathcal{B}$, which will be a subspace of $\mathcal{V}^{\perp}$.

3. Recognition via the Invariant Subspace

We assume that a set of n training data samples from N different entities, or object classes (e.g. images of faces, or whole body appearances), are given, where each class i has n_i samples. Every sample x_j is modeled according to (1), and we concatenate the data into a matrix $X = [X_1, X_2, \cdots, X_N] \in \mathbb{R}^{m \times n}$, where $X_i \in \mathbb{R}^{m \times n_i}$ is the training data matrix obtained by lining up the samples for class i.

Model (1) has been implicitly adopted by the most successful recent approaches to the face recognition problem. In particular, the SRC method [44] aims at "carefully" composing each of the X_i's in such a way that the selected samples are able to represent the salient components s_i's in the best possible way. The matching between a test point $x = s + v$, and a salient component s_i (i.e., the classification), is based on sparse coding and residual computation, and has demonstrated a remarkable robustness against the variation component v, leading to high recognition rates. The SRC approach has been further improved against potential corruptions of the test data point [17, 46]. For instance, [45] improves upon occlusions and computational cost, [32] robustifies the sparse coding problem by computing a sparsity-constrained maximum likelihood solution, [40] simultaneously handles the misalignment, pose and illumination invariance, and [10] addresses the problem of reducing the large amount of training data needed by SRC to be effective.

To address the more general case where also the training data is highly affected by nuisance factors, and a "careful" composition of X is not possible, the SRC approach has been augmented in different ways. In [6] a low-rank matrix recovery [5] approach is designed for pre-processing the corrupted training data. After this step, the SRC method can be applied more effectively. Another approach, [11], proposes to apply sparse coding for modeling the sufficient component by learning a dictionary of prototypes, each of which, given by the average of the data in X_i, is meant to approximate s_i. In addition, sparse coding is also used for modeling the variation subspace. The concatenation of the

[1] A geometric parallel could be observed between the variation and the invariant subspaces with the shared and the private subspaces in the dataset bias problem [23, 37]. However, unlike the shared subspace, used for cross-dataset recognition, the variation space is something we get rid of.

prototype and the variation dictionaries form a new dictionary with which the SRC method can be applied more effectively.

In this work we propose to address the recognition problem with highly corrupted training and testing data by exploiting model (1) in a very different way than previous work. The idea is based on a simple observation. Suppose that the projection operator $P_{\mathcal{V}}$ was available. Then, a test sample x could be processed by computing $x - P_{\mathcal{V}}x = b$. Similarly, for the training dataset, following the property of the invariant subspace, computing $X - P_{\mathcal{V}}X$ produces $[b_1 1_{n_1}^\top, b_2 1_{n_2}^\top, \cdots, b_N 1_{n_N}^\top]$, where b_i is the *invariant component* of class i, and 1_{n_i} is a column vector of ones with length n_i. Therefore, recognition could be done by a simple matching between b and the set of b_i's. This means that corruption (or intra-class variability) in training and testing data, as well as recognition could be handled in a very easy, and efficient way with simple geometry tools.

One major challenge of the proposed approach is posed by the case when two different sufficient components $s_1 \neq s_2$, are such that $s_1 - P_{\mathcal{V}}s_1 = s_2 - P_{\mathcal{V}}s_2$. This means it would be impossible to discriminate between the corresponding classes. The supervised learning approach introduced in the following sections will: (1) allow learning of the invariant subspace, and (2) inherently address the challenge just outlined by promoting a uniform inter-class separation as described in Section 5.2.

4. Invariant Subspace Learning

We begin by observing that since every data point is modeled as $x_j = a_j + b_j$, the training data set X, can be decomposed by $X \doteq A + B$, where $A \in \mathbb{R}^{m\times n}$ collects all the a_j's, and $B \in \mathbb{R}^{m\times n}$ collects all the invariant components, b_j's. We assume that the variation subspace $\mathcal{V}$ has a finite dimension, which is lower than $\min\{m, n\}$. This is reasonable because it states that there are enough data for learning the variation subspace of interest, it allows avoiding overfitting, and it makes the problem tractable. Therefore, attempting to recover A, which in turn allows recovering B, entails solving a low-rank matrix recovery problem.

In practice, the training data will also be affected by noise. We admit that a small percentage of the entries of X are corrupted by values not modeled by the variation and invariant components, which means that such noise should be sparse. This will account for data deviations unlikely to be captured by a finite dimensional linear subspace, such as those induced by image saturations, like image glare, or the presence of strong edges. Therefore, if $E \in \mathbb{R}^{m\times n}$ is the matrix of sparse noise, the model for the training dataset is given by

$$X \doteq A + B + E \, . \qquad (2)$$

Before posing the optimization problem for the estimation of A, and B, we review the standard low-rank matrix recovery problem with sparse noise.

4.1. Low-rank Matrix Recovery

Low-rank (LR) matrix recovery seeks to decompose a data matrix X into $A+E$, where A is a low-rank matrix and E is the associated sparse error. More precisely, given the input data matrix X, LR minimizes the rank of the matrix A while reducing $\|E\|_0$ to derive the low-rank approximation of X. Since the aforementioned optimization problem is NP-hard, [5] proposed to relax the original problem into the following tractable formulation

$$\min_{A,E} \|A\|_* + \alpha\|E\|_1 \qquad \text{s.t.} \;\; X = A + E \, . \qquad (3)$$

In (3), the nuclear norm $\|A\|_*$ (i.e. the sum of the singular values) approximates the rank of A, and the ℓ_0-norm $\|E\|_0$ is replaced by the ℓ_1-norm $\|E\|_1$, which sums up the absolute values of the entries of E. It is shown in [5] that solving the relaxed version of the problem (3) is equivalent to solving the original low-rank matrix approximation problem, as long as the rank of A to be recovered is not too large and the number of errors in E is small (sparse). To solve the optimization problem (3) it is possible to apply the efficient method of augmented Lagrangian multipliers (ALM) [27].

In face recognition X represents the gallery of images of N subjects. By performing the low-rank matrix recovery (3), X gets decomposed into $A = [A_1, \cdots, A_N]$, and $E = [E_1, \cdots, E_N]$. The desired effect is for a subject i to produce a low-rank matrix A_i with columns that look very much alike and span a very narrow space around the sufficient component s_i [6]. The corresponding sparse matrix E_i is expected to pick up the variation components, caused by nuisance factors (e.g., occlusions, disguise, lighting variations, pose, etc.). In [6] the low-rank matrices A_i's are iteratively optimized with robust PCA [5]. In addition, for an increased class separation, a structural incoherence prior is included in the optimization. Other approaches instead, increase discriminability by learning a dictionary in combination with sparse coding and low-rank modeling. In particular, [29] learns a low-rank discriminative dictionary for every class to operate the sparse representation of data samples. [50] instead learns a discriminative dictionary for a sparse and low-rank representation. In [29] testing is similar to the SRC; in [50] the learning of an extra linear multiclass classifier is required.

Unlike previous work, we do not learn a dictionary, and the columns of the low-rank matrix A are meant to span the variation space $\mathcal{V}$, not the space of the sufficient components. Discriminability comes from learning the invariant components B, which leads to a very simple and efficient rule for classification, and can promote class separation with a supervised learning approach described in the following section.

4.2. Supervised Learning

To learn model (2), standard LR (3) is insufficient because we also need to learn the invariant components B. To do so, we need to take into account the geometric, and invariance constraints of (2).

Geometric constraint. In particular, the invariant subspace should be included in the orthogonal complement of the variation subspace $\mathcal{V}^\perp$. Therefore, A and B should satisfy the relationship

$$B^\top A = 0 \; . \tag{4}$$

Invariance constraint. In addition, given two data points $x_1 = a_1 + b_1 + e_1$ and $x_2 = a_2 + b_2 + e_2$, if they are representative of the same class i, the invariant components should be the same, i.e. $b_1 = b_2$. To express this in an algebraic form, b_1 and b_2 should be the solution to the linear system given by the equations $b_1 = \frac{1}{2}(b_1 + b_2)$, and $b_2 = \frac{1}{2}(b_1 + b_2)$. For n data points, where $B = [B_1, B_2, \cdots, B_N]$, the constraint on the invariant components would be $b_1 = b_2 \cdots = b_{n_1}$, for $B_1, \cdots$, and $b_{n-n_N+1} = b_{n-n_N+2} = \cdots = b_n$, for B_N. This can still be expressed in an algebraic form, by generalizing the system of two linear equations to the following expression

$$B(I - Q) = 0 \; , \tag{5}$$

where I is the identity matrix, and Q is a block-diagonal matrix, given by $Q \doteq \mathrm{diag}(\frac{1}{n_1} 1_{n_1} 1_{n_1}^\top, \frac{1}{n_2} 1_{n_2} 1_{n_2}^\top, \cdots, \frac{1}{n_N} 1_{n_N} 1_{n_N}^\top)$, and 1_{n_i} is a column vector with ones of length n_i.

In order to learn A and B, we propose to augment problem (3) with model (2), the *geometric constraint* (4), and the *invariance constraint* (5). In particular, to make the problem more tractable, the geometric and invariance constraints are relaxed to the penalty terms $\|B^\top A\|_F^2$, and $\|B(I - Q)\|_F^2$ in the following optimization problem

$$\begin{aligned} \min_{A,B,E} \; & \|A\|_* + \alpha\|E\|_1 + \beta\|B(I - Q)\|_F^2 + \gamma\|B^\top A\|_F^2 \\ \text{s.t.} \;\; & X = A + B + E \; , \end{aligned} \tag{6}$$

where $\|\cdot\|_F$ indicates the Frobenius norm, and α, β, and γ are penalty weights. Note that the addition of the invariance constraint (5) as a penalty, through Q injects the training dataset labeling information inside the learning problem, turning it into a supervised approach.

4.3. Optimization

Algorithm 1 Invariant Components Learning via the Exact ALM Method

```
Input: Observation matrix $X$, labels $Q$, and penalty weights $\alpha$, $\beta$, $\gamma$
 1: $k = 0;\ \rho > 1;\ \mu_0 > 0;\ \eta = \|X\|_F^2;\ \lambda_0 = \frac{\mathrm{sgn}(X)}{\max(\|\mathrm{sgn}(X)\|_F, \alpha^{-1}\|\mathrm{sgn}(X)\|_\infty)};\ A_0 = 0;\ B_0 = XQ;\ E_0 = 0$
 2: while not converged do
 3:     $j = 0;\ A_k^0 = A_k;\ B_k^0 = B_k;\ E_k^0 = E_k$
 4:     while not converged do
    ▷ Line 5 solves (8)
 5:         $(U, \Sigma, V) = \mathrm{svd}(X - B_k^j - E_k^j + \mu_k^{-1}\lambda_k - \gamma B_k^j B_k^{j\top} A_k^j)$; $A_k^{j+1} = U\mathcal{S}_{(\eta\mu_k)^{-1}}(\Sigma)V^\top$
    ▷ Line 6 solves (9)
 6:         $E_k^{j+1} = \mathcal{S}_{\alpha\mu_k^{-1}}(X - A_k^{j+1} - B_k^{j\top} + \mu_k^{-1}\lambda_k)$
 7:         Update $B_k^{j+1}$ by solving (11) with $A_k^{j+1}$ and $E_k^{j+1}$
 8:         $j \leftarrow j + 1$
 9:     end while
10:     $A_{k+1} = A_k^{j+1}; B_{k+1} = B_k^{j+1}; E_{k+1} = E_k^{j+1}$
11:     $\mu_{k+1} = \rho\mu_k;\ \lambda_{k+1} = \lambda_k + \mu_k(X - A_{k+1} - B_{k+1} - E_{k+1})$
12:     $k \leftarrow k + 1$
13: end while
Output: $A_k, B_k, E_k$
```

In order to solve problem (6), we use the exact ALM method [27], and start by computing the augmented Lagrangian function $L(A, B, E, \lambda)$, given by

$$\begin{aligned} L =& \|A\|_* + \alpha\|E\|_1 + \beta\|B(I - Q)\|_F^2 + \gamma\|B^\top A\|_F^2 \\ & + \langle \lambda, X - A - B - E\rangle + \frac{\mu}{2}\|X - A - B - E\|_F^2 \\ =& \|A\|_* + \alpha\|E\|_1 + \beta\|B(I - Q)\|_F^2 + \gamma\|B^\top A\|_F^2 \\ & + \frac{\mu}{2}\|X - A - B - E + \frac{\lambda}{\mu}\|_F^2 - \frac{1}{2\mu}\|\lambda\|_F^2 \\ =& \|A\|_* + \alpha\|E\|_1 + \beta\|B(I - Q)\|_F^2 + h(A, B, E, \lambda, \mu) \\ & - \frac{1}{2\mu}\|\lambda\|_F^2 \; , \end{aligned} \tag{7}$$

where $\langle X, Y\rangle \doteq \mathrm{trace}(X^\top Y)$, μ is a positive scalar, λ is a Lagrange multiplier matrix, and $h(A, B, E, \lambda, \mu) = \frac{\mu}{2}\|X - A - B - E + \frac{\lambda}{\mu}\|_F^2 + \gamma\|B^\top A\|_F^2$ is a quadratic convenience function. We optimize (7) with an alternating direction strategy, and at every outer iteration of Algorithm 1, A, B, and E are first iteratively updated until convergence; subsequently, λ and μ are updated. The inner iteration updates of Algorithm 1 are given below.

Updating A_{k+1}: From the reduced augmented Lagrangian it is convenient to use the linearization technique of the LADMAP method [28], very effectively used also by other approaches [29, 52, 50], and replace the quadratic term h with its first order approximation, computed at iteration k, and add a proximal term giving the following update

$$\begin{aligned} A_{k+1} =& \arg\min_A \|A\|_* + < \nabla_A h(A_k, B_k, E_k, \lambda_k, \mu_k), \\ & A - A_k > + \frac{\eta\mu_k}{2}\|A - A_k\|_F^2 \\ =& \arg\min_A \|A\|_* + \frac{\eta\mu_k}{2}\|A - (X - B_k - E_k + \frac{\lambda_k}{\mu_k} \\ & - \gamma B_k B_k^\top A_k)\|_F^2 \; , \end{aligned} \tag{8}$$

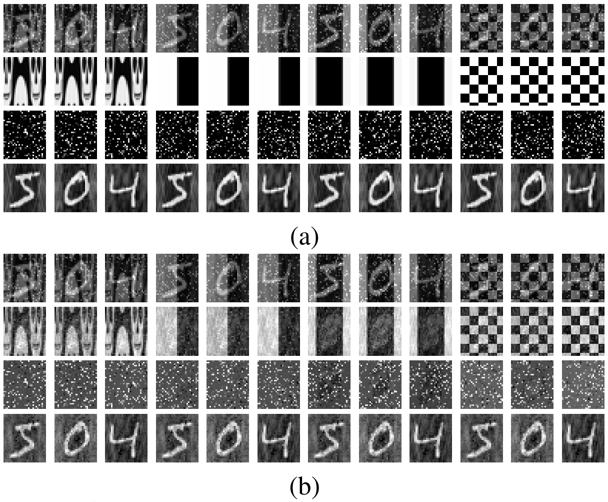

Figure 1. **Synthetic data.** (a) Decomposition of 12 synthetic data points. (b) Decomposition of the same 12 points with Algorithm 1. Top row: input points X. Second row: A components. Third row: Sparse errors E. Bottom row: Invariant components B. Columns with invariant components depicting the same digit belong to the same class. The digits appear "hazy" as a result of being orthogonal to the A components by construction.

where η must be greater than $\|A\|_F^2$ [28]. The solution to (8) is reported in Algorithm 1, and is obtained by applying the singular value thresholding algorithm [3], with the *soft-thresholding shrinkage operator* $\mathcal{S}_\epsilon(x)$, which is equal to: $x - \epsilon$ if $x > \epsilon$, $x + \epsilon$ if $x < -\epsilon$, and 0 elsewhere.

Updating E_{k+1}: From (7), the augmented Lagrangian reduces to

$$E_{k+1} = \arg\min_{E} \alpha\|E\|_1 + \frac{\mu_k}{2}\|E - (X - A_{k+1} - B_k + \frac{\lambda_k}{\mu_k})\|_F^2 \tag{9}$$

and the solution, reported in Algorithm 1, is still obtained with an instance of the singular value thresholding algorithm [3].

Updating B_{k+1}: This update is computed as

$$\begin{aligned} B_{k+1} = \arg\min_{B} \frac{\mu_k}{2}\|X - A_{k+1} - E_{k+1} - B + \frac{\lambda_k}{\mu_k}\|_F^2 + \\ \beta\|B(I-Q)\|_F^2 + \gamma\|B^\top A_{k+1}\|_F^2 \ . \end{aligned} \tag{10}$$

Note that the cost function in (10) is quadratic in B. Therefore, the update can be obtained by computing the partial derivative with respect to B of the cost function, and then setting it to zero. This leads to a Sylvester equation in B, given by

$$\begin{aligned} \gamma A_{k+1}A_{k+1}^\top B + B\left((\beta + \frac{\mu_k}{2})I - 2\beta Q - \beta QQ^\top\right) = \\ \frac{\mu_k}{2}\left(D - A_{k+1} - E_{k+1} + \frac{\lambda_k}{\mu_k}\right) \ . \end{aligned} \tag{11}$$

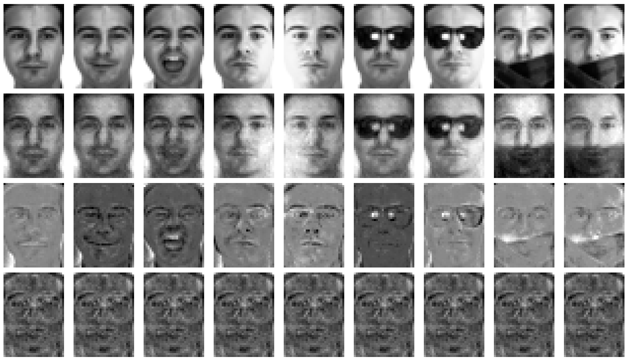

Figure 2. **AR dataset.** Decomposition results for the 13 images of one subject taken in one session. Row meanings are explained in Figure 1. Images are rescaled for better contrast and visualization.

Therefore, the update (10) can be computed with a standard Sylvester equation solver. The full optimization procedure is summarized in Algorithm 1.

5. Classification

Given a test data point x, even if, strictly speaking, we are not in an instance-based learning setting, the obvious approach to perform classification is to compute a label y with a nearest-neighbor (NN) method, where $y = \arg\min_i d(x, B_i)$, and $d(\cdot,\cdot)$ is a suitable distance between x and the invariant matrix B_i, representing class i.

Following the strategy outlined in Section 3, from the invariant components B_i one can estimate $P_{\mathcal{B}_i} : \mathbb{R}^m \rightarrow \mathcal{B}_i$, the operator that projects data points directly onto $\mathcal{B}_i \subset \mathcal{B}$, the invariant subspace for class i. Doing so has the advantage that the projection of x onto $\mathcal{V}^\perp$ gives $b + P_{\mathcal{V}^\perp}e$, whereas the projection of x onto $\mathcal{B}_i$ gives $b + P_{\mathcal{B}_i}e$, and since $\mathcal{B}_i \subset \mathcal{V}^\perp$, it follows that $\|P_{\mathcal{B}_i}e\|_F \leq \|P_{\mathcal{V}^\perp}e\|_F$, which means a lower noise corruption. Therefore, we propose to use the following Frobenius norm $d_F(x, B_i) = n_i^{-\frac{1}{2}}\|B_i - P_{\mathcal{B}_i}x\,1_{n_i}^\top\|_F$. Note that if B_i can be approximated with $b_i 1_{n_i}^\top$, as it normally should, then the distance computation is even faster, because given by

$$d_F(x, B_i) = \|b_i - P_{\mathcal{B}_i}x\|_F \ . \tag{12}$$

5.1. Local Metric Learning

Metric learning improves the performance of the NN classifier if used instead of the Euclidean metric. It has been applied effectively for classification [33], retrieval [19], person reidentification [51, 18], and widely for face verification [8, 15, 34, 35, 20]. Different aspects of metric learning have been investigated, like distance parameters selection, scalability, whether training data should be used in pairs [34], triplets [36] or quadruplets [25], or whether data

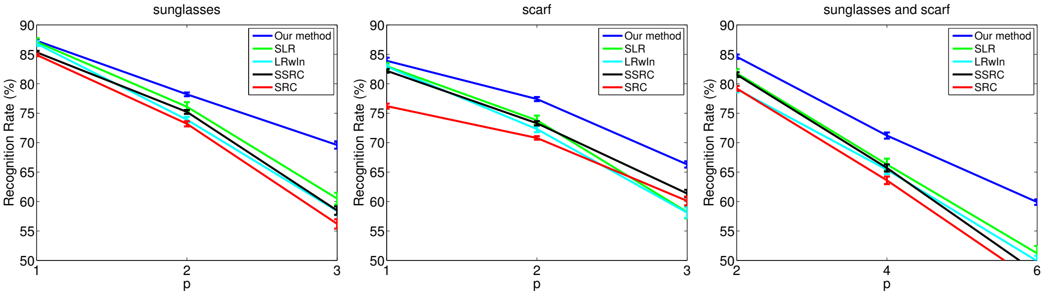

Figure 3. **AR dataset.** Recognition rates versus different numbers p, of corrupted training images per class for the three scenarios: sunglasses (left), scarf (center), sunglasses and scarf (right).

undergoes a linear [9, 24], or nonlinear [4, 20, 38, 47] transformation.

The approach outlined before, which has been derived using geometry, is amenable to an interpretation from a metric learning perspective. Let us recall the definition of Mahalanobis distance between two points x_i and x_j, given by $d_M(x_i, x_j) = \sqrt{(x_i - x_j)^\top M(x_i - x_j)}$, where M is a symmetric positive semi-definite matrix. A *global* linear metric learning method learns a matrix M according to a specific criterion. Since the decomposition $M = L^\top L$ is always possible, the Mahalanobis distance can be expressed also as $d_M(x_i, x_j) = \|L(x_i - x_j)\|_F$.

Global metric learning methods learn the importance and correlation of different input features and take them into account for NN classification, regardless of the specific feature neighborhood where they are applied. Since discriminative power of input features might vary between different neighbors, learning a global metric may be suboptimal. This has motivated the development of *local* metric learning approaches [16, 1, 13, 42, 2], which increase the discriminative power of global Mahalanobis metric learning by learning a number of local metrics.

The proposed approach can be seen as a local metric learning approach, where for the neighborhood of each of the invariant components we learn a Mahalanobis metric. In particular, if $B_i = U_{B_i} S_{B_i} V_{B_i}^\top$ is the singular value decomposition (SVD) of B_i, then the distance (12) can be rewritten as $d_F(x, B_i) = \|U_{B_i} U_{B_i}^\top (x - b_i)\|_F$. This means that $d_F(x, B_i) = d_{M_i}(x, b_i)$, i.e., the Mahalanobis distance between x and b_i, with respect to $M_i = U_{B_i} U_{B_i}^\top$. Therefore, learning a representation based on the invariant components B, is equivalent to learning a set of cluster centers $\{b_i\}$, and a set of Mahalanobis matrices $\{M_i\}$ that act on the neighborhood of each center, and with which labels are assigned based on the NN rule $y = \arg\min_i d_{M_i}(x, b_i)$.

5.2. Class Separation

Most local approaches learn the metrics for each neighborhood independently [42] and require the addition of a form of regularization to avoid overfitting. In contrast, related to [41], our approach learns the metrics jointly, according to the constraints (4) and (5). While the first eliminates the effects of nuisance factors, the second ensures not only invariance, but also class separation. More specifically, since the invariance constraint (5) can be rewritten as $Q = B^\top (BB^\top)^+ B$, it is easy to realize that the Mahalanobis distance $d_M(b_i, b_j)$, with $M = (BB^\top)^+/n$, between the invariant components b_i and b_j, for classes i and j, is such that

$$d_M(b_i, b_j) = \begin{cases} 0 & \text{if } i = j \ , \\ \sqrt{2N} & \text{otherwise} \ , \end{cases} \qquad (13)$$

where for simplicity it is assumed $n_i = n_j$. Without loss of generality, if we assume that the columns of B are zero mean, M is the inverse of the covariance of B (for a short discussion we do not address the rank deficiency of B, which leads to a reduced-rank metric, and to using the pseudoinverse $(BB^\top)^+$). Therefore, (13) means that the invariant subspace $\mathcal{B}$ is such that two different sufficient components s_i and s_j originate two invariant components b_i and b_j that are *different* (i.e., $b_i = s_i - P_{\mathcal{B}} s_i \neq s_j - P_{\mathcal{B}} s_j = b_j$), and *equidistant* (i.e., $d_M(b_i, b_j) = \sqrt{2N}\ \forall i \neq j$), thus promoting a *uniform class separation*.

The observation above suggests also the use of a *global* Mahalanobis metric for NN classification, e.g., in the form of $d_M^2(x, B_i) = n_i^{-1} \sum_{b \in B_i} d_M^2(x, b)$. However, it is more efficient to use the corresponding similarity measure $\kappa_M(b_i, b_j) = b_i^\top (BB^\top)^+ b_j$, which gives 0 if $i \neq j$, and $\frac{1}{n_i}$ if $i = j$. Therefore, we propose the global Mahalanobis similarity measure defined as $\kappa_M(x, B_i) = 1_{n_i}^\top B_i^\top (BB^\top)^+ x$, and the label assignment is done according to $y = \arg\max_i \kappa(x, B_i)$. If $B_i = b_i 1_{n_i}^\top$, the similarity reduces to

$$\kappa_M(x, B_i) = n_i b_i^\top (BB^\top)^+ x \ . \qquad (14)$$

6. Experiments

In order to validate the proposed method we have performed experiments on synthetic data, and on three face recognition datasets. All the results were obtained with a grid search of the parameters α, β, and γ.

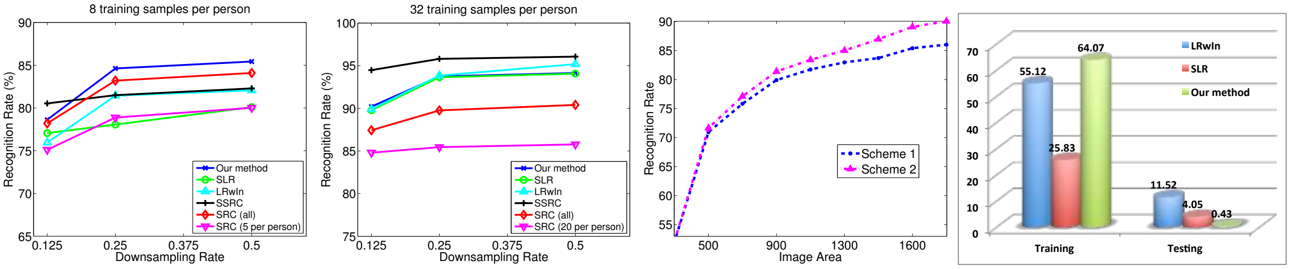

Figure 4. **Extended Yale B dataset.** From left to right: Recognition scores at different image downsampling rates for 8 and 32 training samples per subject; recognition rates obtained with the global metric (14) (Scheme 1) and the local metric (12) (Scheme 2) at various image resolutions and 32 training samples; running time in seconds of our Matlab implementations for training and testing.

Synthetic Data. To empirically verify the convergence of Algorithm 1, we have created a synthetic dataset made of $n = 120$ images of 32×28 pixels, with $N = 10$ invariant components depicting digits, and with image patterns representing A. The synthetic A and B satisfy the constraints (4), and (5), and we have added sparse noise E, corrupting 20% of randomly selected pixels with values drawn from a uniform distribution between 0 and the largest possible pixel value in the image. Figure 1(a) shows the decomposition in A, E, and B of 12 synthetic data points, X (top row), and Figure 1(b) shows the estimated decomposition of the same points. Visually, the recovered decomposition closely resembles the originals, and the coefficients of variation (i.e., $\|\hat{z} - z\|_F / \|z\|_F$ where $\hat{z}$ is the estimated quantity), are 9.71%, 8.67%, 37.2%, for A, B, and E, respectively.

AR Dataset. For this face recognition dataset [31], we follow a protocol used also by other recent works [6, 50]. The dataset contains over 4,000 frontal images of 126 people's faces (70 men and 56 women), images are taken in two sessions and under different facial expressions, illumination conditions and occlusions. In each session 3 images are occluded by sunglasses, 3 by a scarf, and all are taken in different lighting conditions. The images have $165 \times 120 = 19,800$ pixels, and are converted into gray scale, and down-sampled by a factor of 4. As other authors did [6, 50, 11], we select a subset of 50 men and 50 women. Figure 2 illustrates 13 images taken from one subject in one session, along with the decomposition. The proposed algorithm effectively extracts the invariant component (bottom row), which is almost identical for every image, as expected. The second row from the top is a low-rank representation of the face images, and the second row from the bottom is sparse noise. Note how the low-rank representation, a, contains a significant amount of facial features. This is expected because it represents the additive contributions of the variation component, v, plus the projection, $P_{\mathcal{V}} s$, of the sufficient component, s (which is essentially the face), onto the variation subspace, $\mathcal{V}$, which is shared among all the classes.

Following [6, 50] we consider three scenarios, indicated as SUNGLASSES, SCARF and SUNGASSES+SCARF, where we do face recognition with highly corrupted training and testing data. For SUNGLASSES a subject in the training set is composed by p randomly selected face images occluded with sunglasses, and $8 - p$ neutral, all selected from session 1. The remaining $6 - p$ images occluded by sunglasses plus $6 + p$ neutral from both sessions, form 12 testing images per person. Note that face images with sunglasses are occluded about 20%. For the SCARF scenario, the data subdivision is identical only that we consider the face images occluded by a scarf, which produces occlusions of about 40%. For the SUNGLASSES+SCARF case, the difference is that for a given person, p images are occluded with sunglasses and p with the scarf, leaving 17 images for testing per person. Unlike previous work, that have shown results only for $p = 1$, here we also test the case for $p = 2$ and $p = 3$. The experiment has been repeated 5 times and the average recognition rates are plotted in Figure 3. The optimal penalty parameters were $\alpha = 1.5$, $\beta = 1000$, $\gamma = 0.9$. Unless otherwise specified, every result obtained in this section is with the distance (12), i.e., the local metric. Along with ours, we have also tested the structured low-rank representation (SLR) approach [50], the low-rank with incoherence (LRwIn) approach [6], the superposed SRC (SSRC) approach [11], and the SRC [44]. We have reimplemented the SLR, the SSRC, and the LRwIn approaches. For the SRC we have used the code publicly available. Every approach was tested with input images with the same size, and with other parameters set at the peak of their performance. From Figure 3 it can be appreciated that the proposed approach demonstrates a superior robustness with respect to corruption in the training set as p increases. For instance, compared to the overall best competitor, which is SLR, for the SUNGLASSES+SCARF case, for $p = 1$ the improvement is 2.8%, for $p = 2$ is 4.9%, and for $p = 3$ is 8.7%.

Extended-Yale B Dataset. This face recognition dataset [14] contains tightly cropped face images of 38 subjects. Each of them has 59 to 64 images taken under varying lighting conditions, which in total add up to 2,414 images. The cropped images have $192 \times 168 = 32,256$ pixels. We randomly select 8, and in a subsequent experiment 32, training images for each person, and use the rest for testing in a recognition experiment. We repeat this 5 times and report the average recognition rate for the images down-sampled by a factor of 2, 4, and 8. For each of those conditions we also compare against the SLR [50], the LRwIn [6], the SSRC [11], and the SRC [44] approaches at the peak of their performance. For our approach the optimal penalty parameters were $\alpha = 0.9$, $\beta = 1000$, $\gamma = 0.01$. Figure 4 illustrates the comparison between the recognition rates. For the SRC, we also include what happens when the training set drops in size from 8 to 5, and from 32 to 20 training images. This experiment highlights that our approach compares favorably with the others especially when a smaller corrupted training dataset is available, and works on par with others (SRL and LRwIn) with lots of training data. This is because our approach inherently attempts learning a global variation space, shared by all the training data. Even with fewer training images per person their aggregation allows learning the variation space better than in other approaches. SSRC, instead, is the best performer with lots of training data, since it can better learn the variation space for each individual.

Figure 4, right, also shows a comparison between the local metric approach (Scheme 2), based on (12), and the global metric approach (Scheme 1), based on (14), on a subset of the dataset with 32 training data points per person, against different image resolutions. As expected, the local metric learning approach, because it adapts to the invariant component where it operates on, is able to provide better performance. From a geometric perspective, as highlighted in Section 5, the performance drop is justified by the fact that the global approach is not able to filter out as much noise as the local approach is capable of.

Figure 4, far right, shows a running time comparison between the Matlab implementations of ours, the SLR, and the LRwIn methods, running on a high-end PC. Our training procedure appears slightly more costly than the others, but, as anticipated, testing appears faster than SLR by a factor of 10, and faster than LRwIn by a factor of 25.

Dataset	Euclidean	SRC	SSRC	LMNN	Ours
LFW	15.40 ± 0.50	36.91 ± 1.90	46.31 ± 2.43	46.90 ± 1.00	47.10 ± 1.50
AR	29.30 ± 0.50	63.60 ± 0.64	65.70 ± 0.61	62.8 ± 1.00	71.20 ± 0.52

Table 1. **Metric learning.** Comparison between metric learning and other methods on the LFW dataset [22], and on the AR dataset on the scenario SUNGLASSES+SCARF with $p = 2$.

Metric Learning on LFW and AR Datasets. We have tested the large-margin nearest neighbor (LMNN) metric learning approach [43], SRC, SSRC, and ours on the Labeled Faces in the Wild (LFW) dataset [22]. Out of the 13,233 face images of 5749 unique individuals, we selected those with at least 10 images for a total of 143 people and 4174 face images, which were aligned using deep funneling [21], tightly cropped to include only face information, and resized to 106×96 pixels. For each subject, we randomly selected 7 images for training, and the rest were used for testing. The penalty parameters were $\alpha = 0.5$, $\beta = 1000$, $\gamma = 0.2$. The actual processing for both algorithms was repeated 10 times, and was done with the cropped images down-sampled by a factor of 4. In such a scenario with a highly non-linear variation space we obtained the results reported in Table 1, where we also provided results using the baseline Euclidean distance. We also run LMNN and our method on the AR dataset on the highly corrupted scenario given by SUNGLASSES+SCARF with $p = 2$. Table 1 reports the results, which shows that our method performs better especially when robustness against corrupted samples in the gallery is needed.

7. Conclusions

We proposed to represent data by their invariant components. By leveraging recent advances in low-rank matrix recovery, we developed a framework for the supervised learning of invariant components, which corresponds to a metric learning optimization. This representation leads to a simple and efficient testing rule, and promotes inter-class separation. We empirically verified the convergence of the training algorithm, and we applied the model to the face recognition problem with highly corrupted training and testing data. The performance are very promising since they are on par or better than state-of-the-art, with significant gains in time complexity at testing time and in classification accuracy at higher fractions of corrupted training data, as well as with small-size and corrupted training datasets.

References

[1] M. Bilenko, S. Basu, and R. J. Mooney. Integrating constraints and metric learning in semi-supervised clustering. In *ICML*, pages 81–88, 2004. 2, 6

[2] J. Bohne, Y. Ying, S. Gentric, and M. Pontil. Large margin local metric learning. In *ECCV*, volume 8690, pages 679–694, 2014. 6

[3] J. Cai, E. Candés, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010. 5

[4] X. Cai, C. Wang, B. Xiao, X. Chen, and J. Zhou. Deep nonlinear metric learning with independent subspace analysis for face verification. In *ACM Multimedia*, MM '12, pages 749–752, New York, NY, USA, 2012. ACM. 1, 6

[5] E. Candés, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal ACM*, 58(3), 2011. 1, 2, 3

[6] C.-F. Chen, C.-P. Wei, and Y.-C. Wang. Low-rank matrix recovery with structural incoherence for robust face recognition. In *IEEE CVPR*, pages 2618–2625, 2012. 1, 2, 3, 7, 8

[7] B. Cheng, G. Liu, J. Wang, Z. Huang, and S. Yan. Multi-task low-rank affinity pursuit for image segmentation. In *IEEE ICCV*, pages 2439–2446, 2011. 1

[8] Z. Cui, W. Li, D. Xu, S. Shan, and X. Chen. Fusing robust face region descriptors via multiple metric learning for face recognition in the wild. In *IEEE CVPR*, pages 3554–3561, June 2013. 5

[9] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *ICML*, pages 209–216, 2007. 1, 6

[10] W. Deng, J. Hu, and J. Guo. Extended SRC: Undersampled face recognition via intraclass variant dictionary. *IEEE TPAMI*, 34(9):1864–1870, 2012. 2

[11] W. Deng, J. Hu, and J. Guo. In defense of sparsity based face recognition. In *IEEE CVPR*, pages 399–406, 2013. 2, 7, 8

[12] E. Elhamifar and R. Vidal. Robust classification using structured sparse representation. In *IEEE CVPR*, pages 1873–1879, 2011. 1

[13] A. Frome, Y. Singer, and J. Malik. Image retrieval and classification using local distance functions. In *NIPS*, volume 19, pages 417–424, 2007. 2, 6

[14] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: illumination cone models for face recognition under variable lighting and pose. *IEEE TPAMI*, 23(6):643–660, 2001. 7

[15] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? metric learning approaches for face identification. In *IEEE ICCV*, pages 498–505, Sept 2009. 5

[16] T. Hastie and R. Tibshirani. Discriminant adaptive nearest neighbor classification. *IEEE TPAMI*, 18(6):607–616, Jun 1996. 2, 6

[17] R. He, W. S. Zheng, and B. G. Hu. Maximum correntropy criterion for robust face recognition. *IEEE TPAMI*, 33(8):1561–1576, 2011. 2

[18] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *ECCV*, pages 780–793, 2012. 5

[19] S. Hoi, W. Liu, M. Lyu, and W.-Y. Ma. Learning distance metrics with contextual constraints for image retrieval. In *IEEE CVPR*, volume 2, pages 2072–2078, 2006. 5

[20] J. Hu, J. Lu, and Y.-P. Tan. Discriminative deep metric learning for face verification in the wild. In *IEEE CVPR*, pages 1875–1882, June 2014. 5, 6

[21] G. Huang, M. Mattar, H. Lee, and E. G. Learned-Miller. Learning to align from scratch. In *NIPS*, pages 764–772, 2012. 8

[22] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *ECCV*, 2008. 8

[23] A. Khosla, T. Zhou, T. Malisiewicz, A. A. Efros, and A. Torralba. Undoing the damage of dataset bias. In *ECCV*, pages 158–171, 2012. 2

[24] M. Kostinger, M. Hirzer, P. Wohlhart, P. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *IEEE CVPR*, pages 2288–2295, June 2012. 6

[25] M. Law, N. Thome, and M. Cord. Quadruplet-wise image similarity learning. In *IEEE ICCV*, pages 249–256, Dec 2013. 5

[26] J. Lee, B. Shi, Y. Matsushita, I. Kweon, and K. Ikeuchi. Radiometric calibration by transform invariant low-rank structure. In *IEEE CVPR*, pages 2337–2344, 2011. 1

[27] Z. Lin, M. Chen, and Y. Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv:1009.5055*, 2010. 3, 4

[28] Z. Lin, R. Liu, and Z. Su. Linearized alternating direction method with adaptive penalty for low-rank representation. *arXiv*, 2011. 4, 5

[29] L. Ma, C. Wang, B. Xiao, and W. Zhou. Sparse representation for face recognition based on discriminative low-rank dictionary learning. In *IEEE CVPR*, pages 2586–2593, 2012. 3, 4

[30] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Discriminative learned dictionaries for local image analysis. In *IEEE CVPR*, pages 1–8, June 2008. 1

[31] A. M. Martinez and R. Benavente. The AR face database. Technical Report 24, CVC, 1998. 7

[32] Y. Meng, D. Zhang, Y. Jian, and D. Zhang. Robust sparse coding for face recognition. In *IEEE CVPR*, pages 625–632, 2011. 2

[33] T. Mensink, J. Verbeek, F. Perronnin, and G. Csurka. Distance-based image classification: Generalizing to new classes at near-zero cost. *IEEE TPAMI*, 35(11):2624–2637, Nov 2013. 5

[34] A. Mignon and F. Jurie. Pcca: A new approach for distance learning from sparse pairwise constraints. In *IEEE CVPR*, pages 2666–2672, June 2012. 1, 5

[35] H. V. Nguyen and L. Bai. Cosine similarity metric learning for face verification. In *ACCV*, pages 709–720. Springe, 2011. 5

[36] M. Schultz and T. Joachims. Learning a distance metric from relative comparisons. In *NIPS*, 2004. 1, 5

[37] T. Tommasi, N. Quadrianto, B. Caputo, and C. H. Lampert. Beyond dataset bias: Multi-task unaligned shared knowledge transfer. In *ACCV*, 2012. 2

[38] I. W. Tsang, J. T. Kwok, C. Bay, and H. Kong. Distance metric learning with kernels. In *ICANN*, pages 126–129. Citeseer, 2003. 6

[39] R. Vidal and P. Favaro. Low rank subspace clustering (LRSC). *Pattern Recognition Letters*, 43:47–61, 2014. 1

[40] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma. Toward a practical face recognition system: Robust alignment and illumination by sparse representation. *IEEE TPAMI*, 34(2):372–386, 2012. 2

[41] J. Wang, A. Woznica, and A. Kalousis. Parametric local metric learning for nearest neighbor classification. In *NIPS*, pages 1610–1618, 2012. 6

[42] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *JMLR*, 10:207–244, June 2009. 2, 6

[43] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *JMLR*, 10:207–244, 2009. 8

[44] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE TPAMI*, 31(2):210–227, 2009. 1, 2, 7, 8

[45] M. Yang and L. Zhang. Gabor feature based sparse representation for face recognition with gabor occlusion dictionary. In *ECCV*, pages 448–461, 2010. 2

[46] M. Yang, L. Zhang, J. Yang, and D. Zhang. Robust sparse coding for face recognition. In *CVPR*, pages 625–632, 2011. 2

[47] D.-Y. Yeung and H. Chang. A kernel approach for semisupervised metric learning. *IEEE Trans. Neural Networks*, 18(1):141–149, 2007. 6

[48] Q. Zhang and B. Li. Mining discriminative components with low-rank and sparsity constraints for face recognition. In *KDD*, pages 1469–1477, 2012. 1

[49] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja. Low-rank sparse learning for robust visual tracking. In *ECCV*, pages 470–484, 2012. 1

[50] Y. Zhang, Z. Jiang, and L. S. Davis. Learning structured low-rank representations for image classification. In *IEEE CVPR*, pages 676–683, 2013. 1, 3, 4, 7, 8

[51] W.-S. Zheng, S. Gong, and T. Xiang. Reidentification by relative distance comparison. *IEEE TPAMI*, 35(3):653–668, 2013. 5

[52] L. Zhuang, H. Gao, Z. Lin, Y. Ma, X. Zhang, and N. Yu. Non-negative low rank and sparse graph for semi-supervised learning. In *IEEE CVPR*, pages 2328–2335, 2012. 4