Generalized Domain-Adaptive Dictionaries*

Sumit Shekhar Vishal M. Patel Hien V. Nguyen
Rama Chellappa
University of Maryland, College Park, USA
{sshekha,pvishalm,hien,rama}@umiacs.umd.edu

Abstract

Data-driven dictionaries have produced state-of-the-art results in various classification tasks. However, when the target data has a different distribution than the source data, the learned sparse representation may not be optimal. In this paper, we investigate if it is possible to optimally represent both source and target by a common dictionary. Specifically, we describe a technique which jointly learns projections of data in the two domains, and a latent dictionary which can succinctly represent both the domains in the projected low-dimensional space. An efficient optimization technique is presented, which can be easily kernelized and extended to multiple domains. The algorithm is modified to learn a common discriminative dictionary, which can be further used for classification. The proposed approach does not require any explicit correspondence between the source and target domains, and shows good results even when there are only a few labels available in the target domain. Various recognition experiments show that the method performs on par or better than competitive state-of-the-art methods.

1. Introduction

The study of sparse representation of signals and images has attracted tremendous interest in last few years. Sparse representations of signals and images require learning an over-complete set of bases called a dictionary along with linear decomposition of signals and images as a combination of few atoms from the learned dictionary. Olshausen and Field [16] in their seminal work introduced the idea of learning dictionary from data instead of using off-the-shelf bases. Since then, data-driven dictionaries have been shown to work well for both image restoration [3] and classification tasks [26].

The efficiency of dictionaries in these wide range of applications can be attributed to the robust discriminant representations that they provide by adapting to the particular data samples. However, the learned dictionary may not be optimal if the target data has different distribution than the data used for training. These variations are commonplace in vision problems, and can happen due to changes in image sensor (web-cams vs SLRs), camera viewpoint, illumination conditions, etc. It has been shown that such changes can cause significant degradation in classifier performance [2]. Adapting dictionaries to new domains is a challenging task, but has hardly been explored in the vision literature. Yangqing *et al.* [12] considered a special case where corresponding samples from each domain were available, and learned a dictionary for each domain. More recently, Qiu *et al.* [19] proposed a method for adapting dictionaries for smoothly varying domains using regression. However, in practical applications, target domains are scarcely labeled, and domain shifts may result in abrupt feature changes (e.g., changes in resolution when comparing web-cams to DSLRs). Moreover, high dimensional features are often extracted for object recognition. Hence learning a separate dictionary for each domain will have a severe space constraint, rendering it unfeasible for many practical applications.

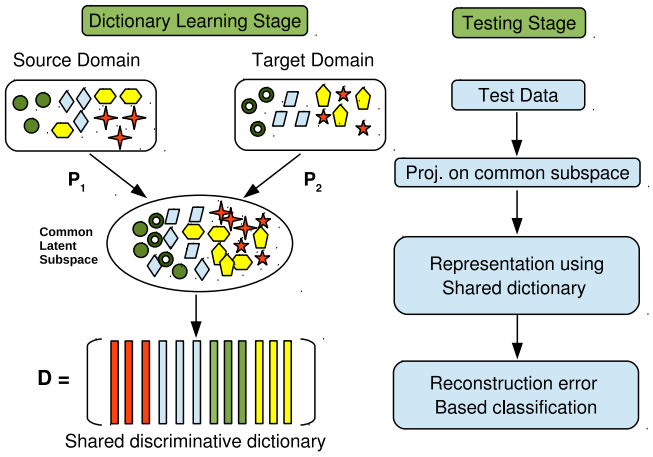


Figure 1. Overview of the proposed dictionary learning method.

*This work was partially supported by an ONR grant N00014-12-1-0124.

In view of the above challenges, we propose a robust method for learning a single dictionary to optimally represent both source and target data. As the features may not be correlated well in the original space, we project data from both the domains onto a common low-dimensional space, while maintaining the manifold structure of data. Simultaneously, we learn a compact dictionary which represents projected data from both the domains well. As the final objective is classification, we learn a class-wise discriminative dictionary. This joint optimization method offers several advantages in terms of generalizability and efficiency of the method. Firstly, learning separate projection matrix for each domain makes it easy to handle any changes in feature dimension and type in different domains. It also makes the algorithm conveniently extensible to handle multiple domains. Further, learning the dictionary on a low-dimensional space makes the algorithm faster, and irrelevant information in original features is discarded. Moreover, joint learning of dictionary and projections ensures that the common internal structure of data in both the domains is extracted, which can be represented well by sparse linear combinations of dictionary atoms.

An additional contribution of the paper is an efficient optimization technique to solve this problem. We will see that by constraining the projection matrices to be orthonormal matrices, convenient forms for optimal dictionary and projection matrices can be obtained. Using the kernel methods, the proposed algorithm can be easily made non-linear, and the resulting optimization problem has a few simple update steps.

1.1. Paper Organization

The paper is organized in six sections. In Section 2, we describe some of the related works. The algorithm is formulated in Section 3, and the optimization technique is described in Section 4. The classification scheme for the learned dictionary is described in Section 5. Experimental results are presented in Section 6, and the final concluding remarks are made in 7.

2. Related Work

The problem of adapting classifiers to new visual domains has recently gained importance in the vision community and several methods have been proposed [21, 13, 6, 5, 11]. Of these methods, Jhuo *et al.* [11] learnt a transformation of source data onto target space, such that the joint representation is low-rank. It however cannot utilize the labeled data while learning the projections. On the other hand, our method jointly learns projections of both the domains, while utilizing the available labels to learn a discriminative dictionary. Han *et al.* [10] suggested learning a shared embedding for different domains, along with a sparsity constraint on the representation. However, they assume pre-learned projections, which may not be optimal. In the dictionary learning literature, Yang *et al.* [27] and Wang *et al.* [24] proposed learning dictionary pairs for cross-modal synthesis. Similarly, methods for joint dimensionality reduction and sparse representation have also been proposed [29, 4, 14, 15]. Additional methods may be found within these references.

3. Problem Framework

The classical dictionary learning approach minimizes the representation error of the given set of data samples subject to a sparsity constraint [1]. Let $\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_N] \in \mathbb{R}^{n \times N}$ be the data matrix. Then, the K-atoms dictionary, $\mathbf{D} \in \mathbb{R}^{n \times K}$, can be trained by solving the following optimization problem

$$\{\mathbf{D}^*, \mathbf{X}^*\} = \arg\min_{\mathbf{D},\mathbf{X}} \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 \text{ s.t. } \|\mathbf{x}_i\|_0 \leq T_0 \ \forall i$$

where, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N] \in \mathbb{R}^{K \times N}$ is the sparse representation of $\mathbf{Y}$ over $\mathbf{D}$, and T_0 is the sparsity level. Here, $\|.\|_0$-norm counts the number of nonzero elements in a vector and $\|.\|_F$ is the Frobenius norm of a matrix.

Now, consider a special case, where we have data from two domains, $\mathbf{Y_1} \in \mathbb{R}^{n_1 \times N_1}$ and $\mathbf{Y_2} \in \mathbb{R}^{n_2 \times N_2}$. We wish to learn a shared K-atoms dictionary, $\mathbf{D} \in \mathbb{R}^{n \times K}$ and mappings $\mathbf{P}_1 \in \mathbb{R}^{n \times n_1}$, $\mathbf{P}_2 \in \mathbb{R}^{n \times n_2}$ onto a common low-dimensional space, which will minimize the representation error in the projected space. Formally, we desire to minimize the following cost function:

$$\mathcal{C}_1(\mathbf{D}, \mathbf{P_1}, \mathbf{P_2}, \mathbf{X_1}, \mathbf{X_2}) = \|\mathbf{P_1}\mathbf{Y_1} - \mathbf{D}\mathbf{X_1}\|_F^2 + \|\mathbf{P_2}\mathbf{Y_2} - \mathbf{D}\mathbf{X_2}\|_F^2$$

subject to sparsity constraints on $\mathbf{X_1}$ and $\mathbf{X_2}$. We further assume that rows of the projection matrices, $\mathbf{P_1}$ and $\mathbf{P_2}$ are orthogonal and normalized to unit-norm. This prevents the solution from becoming degenerate. We will also see that it leads to an efficient scheme for optimization and makes the kernelization of the algorithm possible.

Regularization: It will be desirable if the projections, while bringing the data from two domains to a shared subspace, do not lose too much information available in the original domains. To facilitate this, we add a PCA-like regularization term which preserves energy in the original signal, given as:

$$\mathcal{C}_2(\mathbf{P_1}, \mathbf{P_2}) = \|\mathbf{Y_1} - \mathbf{P_1^T}\mathbf{P_1}\mathbf{Y_1}\|_F^2 + \|\mathbf{Y_2} - \mathbf{P_2^T}\mathbf{P_2}\mathbf{Y_2}\|_F^2.$$

It is easy to show after some algebraic manipulations that the costs $\mathcal{C}_1$ and $\mathcal{C}_2$, after ignoring the constant terms in $\mathbf{Y}$, can be written as:

$$\mathcal{C}_1(\mathbf{D}, \tilde{\mathbf{P}}, \tilde{\mathbf{X}}) = \|\tilde{\mathbf{P}}\tilde{\mathbf{Y}} - \mathbf{D}\tilde{\mathbf{X}}\|_F^2, \quad (1)$$

$$\mathcal{C}_2(\tilde{\mathbf{P}}) = -\text{trace}((\tilde{\mathbf{P}}\tilde{\mathbf{Y}})(\tilde{\mathbf{P}}\tilde{\mathbf{Y}})^T) \quad (2)$$

where,

$$\tilde{\mathbf{P}} = [\mathbf{P_1}\ \mathbf{P_2}],\ \tilde{\mathbf{Y}} = \begin{pmatrix} \mathbf{Y_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{Y_2} \end{pmatrix}, \text{and } \tilde{\mathbf{X}} = [\mathbf{X_1}\ \mathbf{X_2}].$$

Hence, the overall optimization is given as:

$$\{\mathbf{D}^*, \tilde{\mathbf{P}}^*, \tilde{\mathbf{X}}^*\} = \arg\min_{\mathbf{D},\tilde{\mathbf{P}},\tilde{\mathbf{X}}} \mathcal{C}_1(\mathbf{D}, \tilde{\mathbf{P}}, \tilde{\mathbf{X}}) + \lambda \mathcal{C}_2(\tilde{\mathbf{P}})$$
$$\text{s.t. } \mathbf{P_i}\mathbf{P_i^T} = \mathbf{I},\ i = 1, 2 \text{ and } \|\tilde{\mathbf{x}}_j\|_0 \leq T_0, \forall j \quad (3)$$

where, λ is a positive constant.

3.1. Multiple domains

The above formulation can be extended so that it can handle multiple domains. For M domain problem, we simply construct matrices $\tilde{\mathbf{Y}}, \tilde{\mathbf{P}}, \tilde{\mathbf{X}}$ as:

$$\tilde{\mathbf{P}} = [\mathbf{P_1}, \cdots, \mathbf{P_M}],\ \tilde{\mathbf{Y}} = \begin{pmatrix} \mathbf{Y_1} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{Y_M} \end{pmatrix},$$

and

$$\tilde{\mathbf{X}} = [\mathbf{X_1}, \cdots, \mathbf{X_M}].$$

With these definitions, (3) can be extended to multiple domains as follows

$$\{\mathbf{D}^*, \tilde{\mathbf{P}}^*, \tilde{\mathbf{X}}^*\} = \arg\min_{\mathbf{D},\tilde{\mathbf{P}},\tilde{\mathbf{X}}} \mathcal{C}_1(\mathbf{D}, \tilde{\mathbf{P}}, \tilde{\mathbf{X}}) + \lambda \mathcal{C}_2(\tilde{\mathbf{P}})$$
$$\text{s.t. } \mathbf{P_i}\mathbf{P_i^T} = \mathbf{I},\ i = 1, \cdots, M \text{ and } \|\tilde{\mathbf{x}}_j\|_0 \leq T_0, \forall j \quad (4)$$

3.2. Discriminative Dictionary

The dictionary learned in (3) can reconstruct the two domains well, but it cannot discriminate between the data from different classes. Recent advances in learning discriminative dictionaries [20, 28] suggest that learning class-wise, mutually incoherent dictionaries works better for discrimination. To incorporate this into our framework, we write the dictionary $\mathbf{D}$ as $\mathbf{D} = [\mathbf{D_1}, \cdots, \mathbf{D_C}]$, where C is the total number of classes. We modify the cost function similar to [28], which encourages reconstruction samples of a given class by the dictionary of the corresponding class, and penalizes reconstruction by out-of-class dictionaries. The new cost function, $\mathcal{C}_1(\mathbf{D}, \tilde{\mathbf{P}}, \tilde{\mathbf{X}})$ is given as:

$$\mathcal{C}_1(\mathbf{D}, \tilde{\mathbf{P}}, \tilde{\mathbf{X}}) = \|\tilde{\mathbf{P}}\tilde{\mathbf{Y}} - \mathbf{D}\tilde{\mathbf{X}}\|_F^2 + \mu\|\tilde{\mathbf{P}}\tilde{\mathbf{Y}} - \mathbf{D}\tilde{\mathbf{X}}_{\mathbf{in}}\|_F^2 + \nu\|\mathbf{D}\tilde{\mathbf{X}}_{\mathbf{out}}\|_F^2, \quad (5)$$

where μ and ν are the weights given to the discriminative terms, and matrices $\tilde{\mathbf{X}}_{\mathbf{in}}$ and $\tilde{\mathbf{X}}_{\mathbf{out}}$ are given as:

$$\tilde{\mathbf{X}}_{\mathbf{in}}[i,j] = \begin{cases} \tilde{\mathbf{X}}[i,j], & \mathbf{D_i}, \tilde{\mathbf{Y}}_{\mathbf{j}} \in \text{ same class} \\ 0, & \text{otherwise,} \end{cases}$$

$$\tilde{\mathbf{X}}_{\mathbf{out}}[i,j] = \begin{cases} \tilde{\mathbf{X}}[i,j], & \mathbf{D_i}, \tilde{\mathbf{Y}}_{\mathbf{j}} \in \text{ different class} \\ 0, & \text{otherwise.} \end{cases}$$

The cost function is defined only for labeled data in both domains. Unlabeled data can be handled using semi-supervised approaches to dictionary learning [18]. However, we do not explore it further in this paper. Also, note that we do not need to modify the forms of projection matrices, since they capture the overall domain shift, and hence are independent of class variations.

4. Optimization

For the above optimization problem, we can prove the following proposition. The proof is given in the Supplementary Material.

Proposition 1: *There exists an optimal solution* $\mathbf{P_1^*}, \cdots, \mathbf{P_M^*}, \mathbf{D}^*$ *to equation (4), which has the following form:*

$$\mathbf{P_i^*} = (\mathbf{Y_i}\mathbf{A_i})^{\mathbf{T}}\ \forall\ i = 1, \cdots, M \quad (6)$$
$$\mathbf{D}^* = \tilde{\mathbf{P}}^*\tilde{\mathbf{Y}}\tilde{\mathbf{B}} \quad (7)$$

where, $\tilde{\mathbf{P}}^* = [\mathbf{P_1^*}, \cdots, \mathbf{P_M^*}]$, *for some* $\mathbf{A_i} \in \mathbb{R}^{N_i \times n}$ *and some* $\tilde{\mathbf{B}} \in \mathbb{R}^{\sum N_i \times K}$.

With this proposition, the cost functions can be written as:

$$\mathcal{C}_1(\tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\mathbf{X}}) = \|\tilde{\mathbf{A}}^{\mathbf{T}}\tilde{\mathbf{K}}(\mathbf{I} - \tilde{\mathbf{B}}\tilde{\mathbf{X}})\|_F^2 + \mu\|\tilde{\mathbf{A}}^{\mathbf{T}}\tilde{\mathbf{K}}(\mathbf{I} - \tilde{\mathbf{B}}\tilde{\mathbf{X}}_{\mathbf{in}})\|_F^2 + \nu\|\tilde{\mathbf{A}}^{\mathbf{T}}\tilde{\mathbf{K}}\tilde{\mathbf{B}}\tilde{\mathbf{X}}_{\mathbf{out}}\|_F^2 \quad (8)$$
$$\mathcal{C}_2(\tilde{\mathbf{A}}) = -\text{trace}((\tilde{\mathbf{A}}^{\mathbf{T}}\tilde{\mathbf{K}})(\tilde{\mathbf{A}}^{\mathbf{T}}\tilde{\mathbf{K}})^T) \quad (9)$$

where, $\tilde{\mathbf{K}} = \tilde{\mathbf{Y}}^{\mathbf{T}}\tilde{\mathbf{Y}}$ and $\tilde{\mathbf{A}}^{\mathbf{T}} = [\mathbf{A_1^T}, \cdots, \mathbf{A_M^T}]$. The equality constraints now become:

$$\mathbf{P_i}\mathbf{P_i^T} = \mathbf{A_i^T}\mathbf{K_i}\mathbf{A_i} = \mathbf{I},\ \forall i = 1, \cdots, M \quad (10)$$

where, $\mathbf{K_i} = \mathbf{Y_i^T}\mathbf{Y_i}$. The optimization problem now becomes:

$$\{\tilde{\mathbf{A}}^*, \tilde{\mathbf{B}}^*, \tilde{\mathbf{X}}^*\} = \arg\min_{\tilde{\mathbf{A}},\tilde{\mathbf{B}},\tilde{\mathbf{X}}} \mathcal{C}_1(\tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\mathbf{X}}) + \lambda \mathcal{C}_2(\tilde{\mathbf{A}})$$
$$\text{s.t. } \mathbf{A_i^T}\mathbf{K_i}\mathbf{A_i} = \mathbf{I},\ i = 1, \cdots, M \text{ and } \|\tilde{\mathbf{x}}_j\|_1 \leq T_0, \forall j \quad (11)$$

This formulation allows joint update of $\mathbf{D}$ and $\mathbf{P_i}$ via $\mathbf{A_i}$. Also, the form of the cost functions makes it easier to kernelize, which we will see in 4.3.

4.1. Update step for $\tilde{\mathbf{A}}$

Here we assume that $(\tilde{\mathbf{B}}, \tilde{\mathbf{X}})$ are fixed. Then, the optimization for $\tilde{\mathbf{A}}$ can be solved efficiently. We have the following proposition.

Proposition 2: *The optimal solution of equation (11) when* $(\tilde{\mathbf{B}}, \tilde{\mathbf{X}})$ *are fixed is:*

$$\tilde{\mathbf{A}}^* = \mathbf{V}\mathbf{S}^{-\frac{1}{2}}\mathbf{G}^* \tag{12}$$

where, $\mathbf{V}$ *and* $\mathbf{S}$ *come from the eigendecomposition of* $\tilde{\mathbf{K}} = \mathbf{V}\mathbf{S}\mathbf{V}^{\mathbf{T}}$*, and* $\mathbf{G}^* \in \mathbb{R}^{\sum N_i \times n} = [\mathbf{G_1^{*T}}, \cdots, \mathbf{G_M^{*T}}]^T$ *is the optimal solution of the following problem:*

$$\{\mathbf{G}^*\} = \arg\min_{\mathbf{G}} \ \text{trace}[\mathbf{G}^{\mathbf{T}}\mathbf{H}\mathbf{G}]$$
$$s.t.\ \mathbf{G_i^T G_i} = \mathbf{I}\ \forall\ i\ = 1, \cdots, M \tag{13}$$

where,

$$\mathbf{H} = \mathbf{S}^{\frac{1}{2}}\mathbf{V}^{\mathbf{T}}((\mathbf{I} - \tilde{\mathbf{B}}\tilde{\mathbf{X}})(\mathbf{I} - \tilde{\mathbf{B}}\tilde{\mathbf{X}})^{\mathbf{T}} + \mu(\mathbf{I} - \tilde{\mathbf{B}}\tilde{\mathbf{X}}_{\mathbf{in}})$$
$$(\mathbf{I} - \tilde{\mathbf{B}}\tilde{\mathbf{X}}_{\mathbf{in}})^{\mathbf{T}} + \nu(\tilde{\mathbf{B}}\tilde{\mathbf{X}}_{\mathbf{out}})(\tilde{\mathbf{B}}\tilde{\mathbf{X}}_{\mathbf{out}})^{\mathbf{T}} - \lambda\mathbf{I})\mathbf{V}\mathbf{S}^{\frac{1}{2}} \tag{14}$$

Proof: See Supplementary material.

Equation (13) is non-convex due to non-linear equality constraints. Specifically, due to the orthonormality condition on $\mathbf{G_i}$, it involves optimization on the Stiefel manifold. We solved this problem using the efficient approach presented in [25].

4.2. Update step for $\tilde{\mathbf{B}}, \tilde{\mathbf{X}}$

For a fixed $\tilde{\mathbf{A}}$, the problem becomes that of discriminative dictionary learning, with data as $\mathbf{Z} = \tilde{\mathbf{A}}^{\mathbf{T}}\tilde{\mathbf{K}}$ and dictionary $\mathbf{D} = \tilde{\mathbf{A}}^{\mathbf{T}}\tilde{\mathbf{K}}\tilde{\mathbf{B}}$. To jointly learn the dictionary, $\mathbf{D}$, and sparse code, $\tilde{\mathbf{X}}$, we use the framework of the discriminative dictionary learning approach presented in [28]. Once the dictionary, $\mathbf{D}$, is learned, we can update $\tilde{\mathbf{B}}$ as:

$$\tilde{\mathbf{B}} = \mathbf{Z}^{\dagger}\mathbf{D}, \tag{15}$$

where $\mathbf{Z}^{\dagger}$ is the pseudo-inverse of $\mathbf{Z}$ defined as $\mathbf{Z}^{\dagger} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T$.

4.3. Non-linear extension

In many vision problems, projecting the original features may not be good enough due to non-linearity in data. This can be overcome by transforming the data into a high-dimensional feature space. Let $\mathbf{\Phi} : \mathbb{R}^n \rightarrow \mathcal{H}$ be a mapping to the reproducing kernel Hilbert space $\mathcal{H}$. The mapping $\mathcal{P}_\mathbf{i}$ to the reduced space, can ne characterized by a compact, linear operator, $\mathcal{P}_\mathbf{i} : \mathcal{H} \rightarrow \mathbb{R}^d$. Let $\mathcal{K} = \langle\Phi(\tilde{\mathbf{Y}}), \Phi(\tilde{\mathbf{Y}})\rangle_{\mathcal{H}}$. It can be shown similar to proposition 1 that:

$$\mathcal{P}_\mathbf{i}^* = \mathbf{A}^{\mathbf{T}}\mathbf{\Phi}(\mathbf{Y})^{\mathbf{T}}; \mathbf{D}^* = \tilde{\mathbf{A}}^{\mathbf{T}}\mathcal{K}\tilde{\mathbf{B}}.$$

Thus, we get the cost functions as:

$$\mathcal{C}_1(\tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\mathbf{X}}) = \|\tilde{\mathbf{A}}^{\mathbf{T}}\mathcal{K}(\mathbf{I} - \tilde{\mathbf{B}}\tilde{\mathbf{X}}\|_F^2 +$$
$$\mu\|\tilde{\mathbf{A}}^{\mathbf{T}}\mathcal{K}(\mathbf{I} - \tilde{\mathbf{B}}\tilde{\mathbf{X}}_{\mathbf{in}}\|_F^2 + \nu\|\tilde{\mathbf{A}}^{\mathbf{T}}\mathcal{K}\tilde{\mathbf{B}}\tilde{\mathbf{X}}_{\mathbf{out}}\|_F^2, \tag{16}$$
$$\mathcal{C}_2(\tilde{\mathbf{A}}) = -\text{trace}((\tilde{\mathbf{A}}^{\mathbf{T}}\mathcal{K})(\tilde{\mathbf{A}}^{\mathbf{T}}\mathcal{K})^T) \tag{17}$$

and the equality constraints as,

$$\mathbf{A_i^T}\mathcal{K}_\mathbf{i}\mathbf{A_i} = \mathbf{I}\quad \forall\ i = 1, \cdots, M,$$

where $\mathcal{K}_i = \langle\Phi(\mathbf{Y_i}), \Phi(\mathbf{Y_i})\rangle_{\mathcal{H}}$.

5. Classification

Given a test sample, $\mathbf{y_{te}}$ from domain k, we propose the following steps for classification, similar to [15]. We consider the general case of classifying mapping of the sample into kernel space, $\mathbf{\Phi}(\mathbf{y_{te}})$.

1. Compute the embedding of the sample in the common subspace, $\mathbf{z_{te}}$ using the projection, $\mathcal{P}_\mathbf{k}^*$.

   $$\mathbf{z_{te}} = \mathcal{P}_\mathbf{k}^*\mathbf{\Phi}(\mathbf{y_{te}}) = \mathbf{A_k}\mathcal{K}_\mathbf{te}$$

   where, $\mathcal{K}_\mathbf{te} = \langle\mathbf{\Phi}(\mathbf{Y_k}), \mathbf{\Phi}(\mathbf{y_{te}})\rangle$.

2. Compute the sparse coefficients, $\hat{\mathbf{x}}_\mathbf{te}$, of the embedded sample over dictionary $\mathbf{D}$ using the OMP algorithm [17].

   $$\hat{\mathbf{x}}_\mathbf{te} = \arg\min_{\mathbf{x}} \|\mathbf{z_{te}} - \mathbf{D}\mathbf{x}\|_\mathbf{F}^\mathbf{2} \text{ s.t. } \|\mathbf{x}\|_0 \leq T_0.$$

3. Now, the sample can be assigned to class i, if the reconstruction using the class dictionary, $\mathbf{D_i}$ and the sparse code corresponding to the atoms of the dictionary, $\hat{\mathbf{x}}_\mathbf{te}^\mathbf{i}$ is minimum.

   $$\text{Output class} = \arg\min_{i=1,\cdots,C} \|\mathbf{z_{te}} - \mathbf{D_i}\hat{\mathbf{x}}_\mathbf{te}^\mathbf{i}\|_\mathbf{F}^\mathbf{2}.$$

   However, the reconstruction error may not be discriminative enough in the reduced space. So, we project the dictionary, $\mathbf{D_i}$ into the feature space, and assign the test sample to the class with the minimum error in the original feature space:

$$\text{Output class} = \arg\min_{i=1,\cdots,C} \|\mathbf{\Phi}(\mathbf{y_{te}}) - \mathcal{P}_\mathbf{k}^{*\mathbf{T}}\mathbf{D_i}\hat{\mathbf{x}}_\mathbf{te}^\mathbf{i}\|_\mathbf{F}^\mathbf{2}$$
$$= \arg\min_{i=1,\cdots,C} \boldsymbol{\kappa}_\mathbf{te} - \mathbf{2}\mathcal{K}_\mathbf{te}\mathbf{A_k^*D_i} + \hat{\mathbf{x}}_\mathbf{te}^\mathbf{iT}\mathbf{D_i^T}\mathbf{A_k^*}\mathcal{K}_\mathbf{k}\mathbf{A_k^*D_i}\hat{\mathbf{x}}_\mathbf{te}^\mathbf{i},$$

where $\boldsymbol{\kappa}_\mathbf{te} = \langle\mathbf{\Phi}(\mathbf{y_{te}}), \mathbf{\Phi}(\mathbf{y_{te}})\rangle$. The proposed, Shared Domain-adapted Dictionary Learning (SDDL) algorithm is summarized in Algorithm 1.

6. Experiments

We conducted various experiments to ascertain the effectiveness of the proposed method. First, we demonstrate some synthesis and recognition results on the CMU Multi-Pie dataset for face recognition across pose and illumination variations. This also provides insights into our method through visual examples. Next we show the performance of our method on domain adaptation databases and compare it with existing adaptation algorithms.

Input: Data $\{\mathbf{Y_i}\}_{\mathbf{i}=1}^{M}$ and corresponding class labels $\{C_i\}_{i=1}^{M}$ for M domains, sparsity level T_0, dictionary size K and dimension n, parameter values μ, ν

Procedure:

1. *Initialize:* Initialize $\tilde{\mathbf{A}}$ such that $\mathbf{A_i}\boldsymbol{\mathcal{K}}_\mathbf{i}\mathbf{A_i} = \mathbf{I}$ $\forall\, i = 1, \cdots, M$. For this, find SVD of each kernel matrix, $\boldsymbol{\mathcal{K}}_\mathbf{i} = \mathbf{V_i S_i V_i^T}$. Set $\mathbf{A}_i$ as the matrix of eigen-vectors with top n eigen-values as columns.

2. *Update step for* $\tilde{\mathbf{B}}$*:* Learn common dictionary $\mathbf{D}$ with data as $\mathbf{Z} = \tilde{\mathbf{A}}^\mathbf{T}\boldsymbol{\mathcal{K}}$, and using discriminative dictionary learning algorithm as FDDL. Update $\tilde{\mathbf{B}}$ as:

$$\tilde{\mathbf{B}} = \mathbf{Z}^\dagger\mathbf{D}$$

3. *Update step for* $\tilde{\mathbf{A}}$*:* Update $\tilde{\mathbf{A}}$ as:

$$\{\mathbf{G}^*\} = \arg\min_{\mathbf{G}} \text{ trace}[\mathbf{G^T H G}]$$
$$s.t.\ \mathbf{G_i^T G_i} = \mathbf{I}\,\forall\, i\ = 1, \cdots, M$$

where, $\tilde{\mathbf{A}}^* = \mathbf{V S}^{-\frac{1}{2}}\mathbf{G}^*$ and $\mathbf{H}$ is:

$$\mathbf{H} = \mathbf{S}^{\frac{1}{2}}\mathbf{V}^\mathbf{T}((\mathbf{I} - \tilde{\mathbf{B}}\tilde{\mathbf{X}})(\mathbf{I} - \tilde{\mathbf{B}}\tilde{\mathbf{X}})^\mathbf{T} + \mu(\mathbf{I} - \tilde{\mathbf{B}}\tilde{\mathbf{X}}_\mathbf{in})$$
$$(\mathbf{I} - \tilde{\mathbf{B}}\tilde{\mathbf{X}}_\mathbf{in})^\mathbf{T} + \nu(\tilde{\mathbf{B}}\tilde{\mathbf{X}}_\mathbf{out})(\tilde{\mathbf{B}}\tilde{\mathbf{X}}_\mathbf{out})^\mathbf{T} - \lambda\mathbf{I})\mathbf{V}\mathbf{S}^{\frac{1}{2}}$$

Output: Learned dictionary $\mathbf{D}$, projection matrices $\{\mathbf{A_i}\}_{\mathbf{i}=1}^{M}$

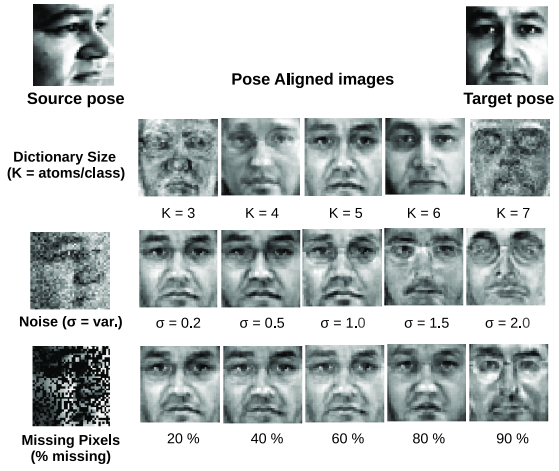
Algorithm 1: Shared Domain-adapted Dictionary Learning (SDDL)

6.1. CMU Multi-Pie Dataset

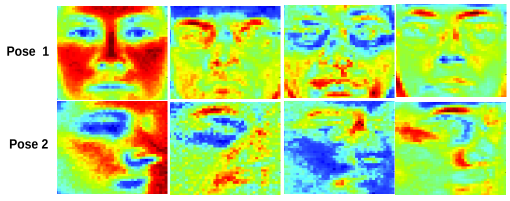
The Multi-pie dataset [9] is a comprehensive face dataset of 337 subjects, having images taken across 15 poses, 20 illuminations, 6 expressions and 4 different sessions. For the purpose of our experiment, we used 129 subjects common to both Session 1 and 2. The experiment was done on 5 poses, ranging from frontal to 75°. Frontal faces were taken as the source domain, while different off-frontal poses were taken as target domains. Dictionaries were trained using illuminations $\{1, 4, 7, 12, 17\}$ from the source and the target poses, in Session 1 per subject. All the illumination images from Session 2, for the target pose, were taken as probe images. The linear kernel was used for all the experiments.

6.1.1 Pose Alignment

First we consider the problem of pose alignment using the proposed dictionary learning framework. Pose alignment is challenging due to the highly non-linear changes induced by 3-D rotation of face. Images at the extreme pose of 60° were taken as the target pose. A shared discriminative dictionary was learned using the approach described in this paper. Given the probe image, it was projected on the latent subspace and reconstructed using the dictionary. The reconstruction was back-projected onto the source pose domain, to give the aligned image. Figure 2(a) shows the synthesized images for various conditions. We can draw some useful insights about the method from this figure. Firstly, it can be seen that there is an optimal dictionary size, $K = 5$, where the best alignment is achieved. Further, by learning a discriminative dictionary, the identity of the subject is retained. For $K = 7$, the alignment is not good, as the learned dictionary is not able to successfully correlate the two domains when there are more atoms in the dictionary. Dictionary with $K = 3$ has higher reconstruction error, hence the result is not optimal. We chose $K = 5$ for additional experiments with noisy images. It can be seen that from rows 2 and 3 that the proposed method is robust even at high levels of noise and missing pixels. Moreover, de-noised and inpainted synthesized images are produced as shown in rows 2 and 3 of Figure 2(a), respectively. This shows the effectiveness of our method. Moreover, the learned projection matrices (Figure 2(b)) show that our method can learn the internal structure of the two domains. As a result, it is able to learn a robust common dictionary.



(a)



(b)

Figure 2. (a) Examples of pose-aligned images using the proposed method. Synthesis in various conditions demonstrate the robustness of the method. (b) First few components of the learned projection matrices for the two poses.

6.1.2 Recognition

We also conducted recognition experiment using the set-up described above. Table 1 shows that our method compares favorably with some of the recently proposed multi-view recognition algorithms [23], and gives the best performance on average. The dictionary learning algorithm, FDDL [28] is not optimal here as it is not able to efficiently represent the non-linear changes introduced by the pose variation.

Method	Probe pose					Average
	15°	30°	45°	60°	75°	
PCA	15.3	5.3	6.5	3.6	2.6	6.7
PLS [22]	39.3	40.5	41.6	41.1	38.7	40.2
LDA	98.0	94.2	91.7	84.9	79.0	89.5
CCA [22]	92.1	89.7	88.0	86.1	83.0	83.5
GMLDA [23]	**99.7**	**99.2**	98.6	94.9	95.4	97.6
FDDL [28]	96.8	90.6	94.4	91.4	90.5	92.7
SDDL	98.4	98.2	**98.9**	**99.1**	**98.8**	**98.7**

Table 1. Comparison of the proposed method with other algorithms for face recognition across pose.

6.2. Object Recognition

We now evaluate our method for object recognition. Performance of the proposed SDDL method is compared to FDDL [28], and some recently proposed domain-adaptation algorithms [21, 6, 5, 11].

6.2.1 Experimental Set-up

The experiments use the dataset which was introduced in [21]. The dataset consists of images from 3 sources: Amazon (consumer images from online merchant sites), DSLR (images by DSLR camera) and Webcam (low quality images from webcams). In addition, we also tested on the Caltech-256 dataset [8], taking it as the fourth domain. Figure 3 shows sample images from these datasets, and clearly highlights the differences between them. We follow 2 set-ups for testing the algorithm. In the first set-up, 10 common classes: BACKPACK, TOURING-BIKE, CALCULATOR, HEADPHONES, COMPUTER- KEYBOARD, LAPTOP-101, COMPUTER- MONITOR, COMPUTER-MOUSE, COFFEE-MUG, AND VIDEO- PROJECTOR, common to all the four datasets are used. In this case, there are a total of 2533 images. Each category has 8 to 151 images in a dataset. In the second set-up, we evaluate the methods for adaptation using multiple domains. In this case, we restrict to the first dataset, and test on all the 31 classes in it. For both the cases, we use 20 training samples per class for Amazon/Caltech, and 8 samples per class for DSLR/Webcam when used as source, and 3 training samples for all of them when used for target domain. Rest of the data in the target domain is used for testing. The experiment is run 20 times for random train/test splits and the result is averaged over all the runs.

Figure 3. Example images from KEYBOARD and BACK-PACK categories in Caltech-256, Amazon, Webcam and DSLR. Caltech-256 and Amazon datasets have diverse images, Webcam and DSLR are similar datasets with mostly images from offices.

Feature Extraction: We used the 800-bin SURF features provided by [21] for the Amazon, DSLR and Webcam datasets. For the Caltech images, first SURF features were extracted from the images of the Caltech data and a random subset of the Amazon dataset. The features obtained from the Amazon dataset were grouped into 800 clusters using the k-means algorithm. The cluster centers were then used to quantize the SURF features obtained from the Caltech data to form 800-bin histograms. The histograms were normalized and then used for classification.

Parameter Settings: For our SDDL method, we used the simple non-parametric histogram intersection kernel for reporting all the values. We set $\mu = 4$ and $\nu = 30$. Dictionary size, $K = 4$ atoms per class and final dimension, $n = 60$ for the first set-up. For the second set-up, $K = 6$ atoms per class and $n = 90$. For FDDL, the parameters, μ and ν are the same as SDDL, and we learn $K = 8$ atoms per class for the first set-up and $K = 10$ atoms per class for the second. The FDDL dictionary was trained using both the source and the target domain features, as it was found to give the best results. Original histogram features were used for both the algorithms.

6.2.2 Results using single source

Table 2(a) shows a comparison of the results of different methods on 8 source-target pairs. The proposed algorithm gives the best performance for 5 domain pairs, and is the second best for 2 pairs. For Caltech-DSLR and Amazon-Webcam domain pairs, there is more than 15% improvement over the GFK algorithm [5]. Furthermore, a com-

(a) Performance comparison on single source four domains benchmark (C: caltech, A: amazon, D: dslr, W: webcam)

Methods	C → A	C → D	A → C	A → W	W → C	W → A	D → A	D → W
Metric[21]	33.7 ± 0.8	35.0 ± 1.1	27.3 ± 0.7	36.0 ± 1.0	21.7 ± 0.5	32.3 ± 0.8	30.3 ± 0.8	55.6 ± 0.7
SGF[6]	40.2 ± 0.7	36.6 ± 0.8	37.7 ± 0.5	37.9 ± 0.7	29.2 ± 0.7	38.2 ± 0.6	39.2 ± 0.7	69.5 ± 0.9
GFK[5]	46.1 ± 0.6	55.0 ± 0.9	$\mathbf{39.6 \pm 0.4}$	56.9 ± 1.0	$\mathbf{32.8 \pm 0.1}$	46.2 ± 0.6	46.2 ± 0.6	$\mathbf{80.2 \pm 0.4}$
FDDL[28]	39.3 ± 2.9	55.0 ± 2.8	24.3 ± 2.2	50.4 ± 3.5	22.9 ± 2.6	41.1 ± 2.6	36.7 ± 2.5	65.9 ± 4.9
SDDL	$\mathbf{49.5 \pm 2.6}$	$\mathbf{76.7 \pm 3.9}$	27.4 ± 2.4	$\mathbf{72.0 \pm 4.8}$	29.7 ± 1.9	$\mathbf{49.4 \pm 2.1}$	$\mathbf{48.9 \pm 3.8}$	72.6 ± 2.1

(b) Performance comparison on multiple sources three domains benchmark

Source	Target	SGF* [7]	SGF [6]	RDALR[11]	FDDL[28]	SDDL
dslr, amazon	webcam	$\mathbf{64.5 \pm 0.3}$	52 ± 2.5	36.9 ± 1.1	41.0 ± 2.4	57.8 ± 2.4
amazon, webcam	dslr	51.3 ± 0.7	39 ± 1.1	31.2 ± 1.3	38.4 ± 3.4	$\mathbf{56.7 \pm 2.3}$
webcam, dslr	amazon	$\mathbf{38.4 \pm 1.0}$	28 ± 0.8	20.9 ± 0.9	19.0 ± 1.2	24.1 ± 1.6

Table 2. Comparison of the performance of the proposed method on the Amazon, Webcam, DSLR and Caltech datasets. SGF* [7] refers to the PAMI submission of the author, currently under review. Numbers obtained by personal communication.

parison with the FDDL algorithm shows that the learning framework of [28] is inefficient, when the test data comes from a different distribution than the data used for training.

6.2.3 Results using multiple sources

As our proposed framework can also handle multiple domains, we also experimented with multiple source adaptation. Table 2 (b) shows the results for 3 possible combinations. Our method outperforms the original SGF method [6] on two settings, and other methods for all the settings. However, [7] reports higher numbers on webcam and amazon as targets, using boosted classifiers. Similarly techniques can be explored for improving the proposed method as a future direction.

6.2.4 Ease of adaptation

A rank of domain (ROD) metric was introduced in [5] to measure the adaptability of different domains. It was shown that ROD correlates with the performance of adaptation algorithm. For example, Amazon-Webcam pair has higher ROD than DSLR-Webcam pair, hence, GFK performs worse on the former. However, for our case, we find that the recognition rates for these cases are 72.0 % and 72.6 %, respectively. This is the case because by learning projections along-with the common dictionary, we can achieve a better alignment of the datasets.

6.2.5 Parameter Variations

We also conducted experiments studying recognition performance under different input parameters. Figure 4 shows the result of different settings. The implications are briefly discussed below:

1. **Number of source images:** Here, we choose Amazon/Webcam domain pair, as it is "difficult" to adapt. We increased the number of source images and studied the performance of SDDL and compared it with FDDL. It can be seen that while FDDL's performance decreases sharply with more source images, SDDL method shows an increase in the performance. Hence, by adapting the source to the target domain, our method can use the source information to increase the accuracy of target recognition, even when their distributions are very different.

2. **Dictionary size:** All the domain pairs show an initial sharp increase in the performance, and then become almost flat after the dictionary size of 3 or 4. The flat region indicates that alignment of the source and the target data is limited by the number of available target samples. But also, on a positive note, it can be seen that even a smaller dictionary can give the optimal performance.

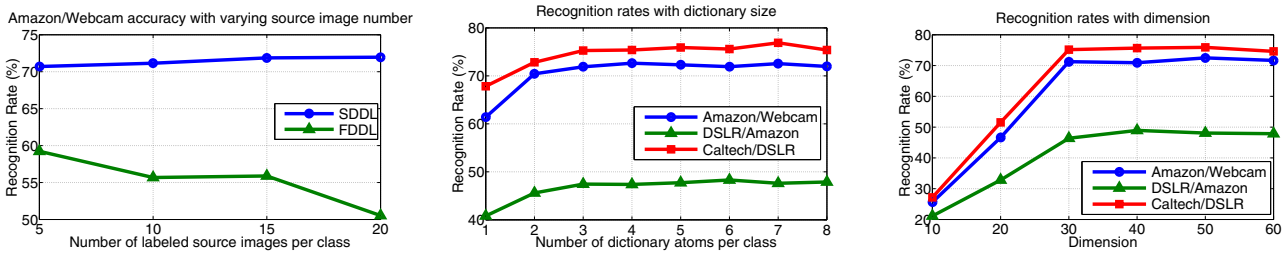
3. **Common subspace dimension:** Similar to the previous case, we get an initial sharp increase followed by a flat recognition curve. This shows that the method is effective even when the data is projected onto a low-dimensional space.

7. Conclusion

We have proposed a novel framework for adapting dictionaries to testing domains under arbitrary domain shifts. An efficient optimization method is presented. Furthermore, the method is kernelized so that it is robust and can deal with the non-linearity present in the data. The learned dictionary is compact and low-dimensional. We show that the method achieves the state-of-the-art performance on the object and face recognition databases. Future works will include studying the effect of using unlabeled data while training, and other relevant problems like large-scale and online adaptation of dictionaries.

References

[1] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE*



(a) (b) (c)

Figure 4. Recognition performance under different: (a) number of source images, (b) dictionary size, and (c) common subspace dimension. Naming of domains is done as source/target.

Transactions on Signal Processing, 54(11):4311–4322, Nov 2006. 2

[2] H. Daumé III. Frustratingly easy domain adaptation. In *ACL*, 2007. 1

[3] M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 15(12):3736–3745, Dec 2006. 1

[4] I. Gkioulekas and T. Zickler. Dimensionality reduction using the sparse linear model. In *NIPS*, pages 271–279, 2011. 2

[5] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2066–2073, June 2012. 2, 6, 7

[6] R. Gopalan, R. Li, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *IEEE International Conference on Computer Vision*, 2011. 2, 6, 7

[7] R. Gopalan, R. Li, and R. Chellappa. Unsupervised adaptation across domain shift by generating intermediate data representations. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Under Review. 7

[8] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007. 6

[9] R. Gross, I. Matthews, J. F. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image Vision Computing*, 28(5):807–813, 2010. 5

[10] Y. Han, F. Wu, D. Tao, J. Shao, Y. Zhuang, and J. Jiang. Sparse unsupervised dimensionality reduction for multiple view data. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(10):1485–1496, Oct. 2012. 2

[11] I.-H. Jhuo, D. Liu, D. Lee, and S.-F. Chang. Robust visual domain adaptation with low-rank reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2168–2175, June 2012. 2, 6, 7

[12] Y. Jia, M. Salzmann, and T. Darrell. Factorized latent spaces with structured sparsity. In *NIPS*, pages 982–990, 2010. 1

[13] B. Kulis, K. Saenko, and T. Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1785–1792, June 2011. 2

[14] J. Mairal, F. Bach, and J. Ponce. Task-driven dictionary learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):791–804, April 2012. 2

[15] H. V. Nguyen, V. M. Patel, N. M. Nasrabadi, and R. Chellappa. Sparse embedding: A framework for sparsity promoting dimensionality reduction. In *European Conference on Computer Vision*, pages 414–427, Oct. 2012. 2, 4

[16] B. Olshausen and D. Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997. 1

[17] Y. Pati, R. Rezaiifar, and P. Krishnaprasad. Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. In *Asilomar Conference on Signals, Systems and Computers*, 1993. 4

[18] J. Pillai, A. Shrivastava, V. Patel, and R. Chellappa. Learning discriminative dictionaries with partially labeled data. In *IEEE Conference on Image Processing*, Oct. 2012. 3

[19] Q. Qiu, V. M. Patel, P. Turaga, and R. Chellappa. Domain adaptive dictionary learning. In *ECCV*, 2012. 1

[20] I. Ramírez, P. Sprechmann, and G. Sapiro. Classification and clustering via dictionary learning with structured incoherence and shared features. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3501–3508, June 2010. 3

[21] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *ECCV*, volume 6314, pages 213–226, 2010. 2, 6, 7

[22] A. Sharma and D. W. Jacobs. Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 593–600, June 2011. 6

[23] A. Sharma, A. Kumar, H. D. III, and D. W. Jacobs. Generalized multiview analysis: A discriminative latent space. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2160–2167, June 2012. 6

[24] S. Wang, L. Zhang, Y. Liang, and Q. Pan. Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2216–2223, June 2012. 2

[25] Z. Wen and W. Yin. A feasible method for optimization with orthogonality constraints. Technical report, Rice University, 2010. 4

[26] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227, 2009. 1

[27] J. Yang, Z. Wang, Z. Lin, S. Cohen, and T. Huang. Coupled dictionary training for image super-resolution. *IEEE Transactions on Image Processing*, 21(8):3467–3478, Aug. 2012. 2

[28] M. Yang, L. Zhang, X. Feng, and D. Zhang. Fisher Discrimination Dictionary learning for sparse representation. In *IEEE International Conference on Computer Vision*, pages 543–550, Nov. 2011. 3, 4, 6, 7

[29] L. Zhang, M. Yang, Z. Feng, and D. Zhang. On the dimensionality reduction for sparse representation based face recognition. In *International Conference on Pattern Recognition*, pages 1237–1240, Aug. 2010. 2