Coupled Dictionary and Feature Space Learning with Applications to Cross-Domain Image Synthesis and Recognition

De-An Huang and Yu-Chiang Frank Wang
Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
andrew800619@gmail.com, ycwang@citi.sinica.edu.tw

Abstract

Cross-domain image synthesis and recognition are typically considered as two distinct tasks in the areas of computer vision and pattern recognition. Therefore, it is not clear whether approaches addressing one task can be easily generalized or extended for solving the other. In this paper, we propose a unified model for coupled dictionary and feature space learning. The proposed learning model not only observes a common feature space for associating cross-domain image data for recognition purposes, the derived feature space is able to jointly update the dictionaries in each image domain for improved representation. This is why our method can be applied to both cross-domain image synthesis and recognition problems. Experiments on a variety of synthesis and recognition tasks such as single image super-resolution, cross-view action recognition, and sketch-to-photo face recognition would verify the effectiveness of our proposed learning model.

1. Introduction

Many computer vision problems can be approached as solving the task of associating data or knowledge across different domains. For example, as depicted in Figure 1, image super-resolution (SR) [5] takes one or multiple low-resolution (LR) images for producing the corresponding high-resolution (HR) versions. On the other hand, cross-view action recognition utilizes training data captured by one camera, and thus the designed features or classifiers can be applied to recognize test data at a different view [4]. For the above cross-domain image *synthesis* (e.g., image SR) and *recognition* (e.g., cross-view action recognition) problems, how to represent and relate data across different domains become a major challenge [20, 25, 10, 16, 12].

With the goal to transfer the knowledge from the source to target domain, recent developments in *transfer learning* [15] have shown promising results for cross-domain recognition problems. Among techniques for addressing such recognition tasks, *domain adaptation* [1] particularly favors the scenarios in which labeled data can be obtained at the source domain, but only little or no labeled target domain data is available. As a result, *unlabeled* data from both domains will be utilized for relating the knowledge across different domains. Generally, approaches like [12, 16, 18, 10] focus on determining a *common feature space* or *representation* using cross-domain unlabeled data pairs, so that classifiers trained in this feature space can be applied to recognize the projected test data. For example, Li *et al.* [10] determined a feature subspace via canonical correlation analysis (CCA) [8] for recognizing faces with different poses. For cross-camera action recognition, Liu *et al.* [12] proposed a bag-of-bilingual-words (BoBW) model as a shared feature representation, which is used to describe the same action data captured by different cameras. A Partial Least Squares (PLS) based framework was recently proposed by Sharma and Jacobs [16] for solving cross-domain image recognition. As pointed out in [19], although the above feature spaces well preserve cross-domain data structures (e.g., data correlation), they cannot be easily extended to image synthesis problems due to the lack of data representation or reconstruction guarantees.

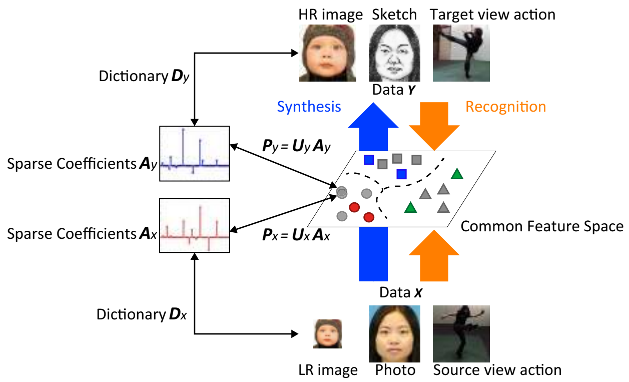

Figure 1. Illustration of cross-domain image synthesis or recognition problems. Note that $\mathbf{D}$, $\mathbf{A}$, $\mathbf{U}$, and $\mathbf{P}$ are the dictionaries, coefficients, projection matrices, and projected data observed at the associated image domain (i.e., data $\mathbf{X}$ or $\mathbf{Y}$), respectively.

For image synthesis, one typically deals with raw or noisy input data for recovering its desirable version. Among existing approaches, *coupled dictionary learning* assumes that some relationships between raw and desirable image data exist and aims at learning a pair of dictionaries for describing cross-domain image data. As a result, information extracted from the input domain can be applied to synthesize images at the output domain accordingly. For example, Yang *et al.* [25] assumed that LR image patches have the *same* sparse representations as their HR versions do, and proposed a *joint dictionary learning* model for SR using concatenated HR/LR image features. They later imposed relaxed constraints on the observed dictionary/coefficient pairs across image domains for improved performance [24]. Wang *et al.* [19] further proposed a semi-coupled dictionary learning (SCDL) scheme by advancing a linear mapping for cross-domain image sparse representation. Their method has been successfully applied to applications of image SR and cross-style synthesis.

In addition to the aforementioned assumptions on image priors, most prior image synthesis algorithms focused on data representation/reconsturction when designing or optimizing their proposed formulation. As argued in [7], if one needs to perform classification after obtaining the desirable output images (e.g., face recognition after hallucination), it would be preferable to integrate image synthesis and recognition algorithms into a unified framework instead of solving them separately. Another potential yet practical issue of the most prior synthesis approaches is that, their need to collect cross-domain training image data beforehand might not be applicable for real-world applications like single image SR or denoising.

It is worth noting that, sparse representation has been widely applied to various image synthesis and recognition tasks [3, 25, 22]. Besides the aforementioned work of image SR [25], Elad and Aharon [3] proposed to utilize an overcomplete dictionary observed from an input noisy image, and thus the associated noise patterns can be removed from the reconstructed image for denoising purposes. The formulation of sparse representation was also applied by Wright *et al.* for recognizing face images [22]. Recently, Zhang *et al.* [26] addressed both face restoration and recognition problems by jointly estimating the blurring kernel and sparse representation. As noted in [16, 12], however, the use of a single linear operator for relating face images and their degraded versions might not be preferable for general image recognition problems. Nevertheless, sparse representation has been shown to be a very effective technique in representing or recognizing image data.

1.1. Our Contributions

The main contribution of this paper is to present a joint model which learns a pair of dictionaries with a feature space for describing and associating cross-domain data. Since our proposed model iterates between the stages of coupled dictionary and feature space learning during optimization, we not only learn a common feature space for relating cross-domain image data, this derived feature space will be utilized to update the observed dictionary pair for improved data representation in each domain. Therefore, our model is able to address both cross-domain synthesis and recognition problems, while most existing works (e.g., [16, 19]) focus on solving *either* task and lack the ability for the other. As confirmed later by our experiments, our proposed model can be applied to a variety of cross-domain image synthesis and recognition tasks such as *single image super-resolution, cross-camera action recognition*, and *sketch-to-photo face recognition*.

2. Coupled Dictionary and Feature Space Learning

In Section 2.1, we present the problem formulation and explain how we represent and associate cross-domain image data by jointly solving coupled dictionary and common feature space learning problems. Optimization details for the training stage of our model are presented in Section 2.2.

2.1. Problem Formulation

Let image sets $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbf{R}^{d_1 \times n}$ and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n] \in \mathbf{R}^{d_2 \times n}$ be n unlabeled data pairs extracted from two different domains, whose dimensions are d_1 and d_2, respectively. Coupled dictionary learning can be approached as solving the following minimization problem:

$$\begin{aligned}\min_{\mathbf{D}_x, \mathbf{D}_y, \mathbf{A}_x, \mathbf{A}_y} \; & E_{DL}(\mathbf{X}, \mathbf{D}_x, \mathbf{A}_x) + E_{DL}(\mathbf{Y}, \mathbf{D}_y, \mathbf{A}_y) \\ & + E_{Coupled}(\mathbf{D}_x, \mathbf{D}_y, \mathbf{A}_x, \mathbf{A}_y).\end{aligned} \tag{1}$$

In (1), E_{DL} denotes the energy term for dictionary learning and is typically in terms of data reconstruction error. The coupled energy term $E_{Coupled}$ regularizes the relationship between the observed dictionaries $\mathbf{D}_x \in \mathbf{R}^{d_x \times k_1}$ and $\mathbf{D}_y \in \mathbf{R}^{d_y \times k_2}$, or that between the resulting coefficients $\mathbf{A}_x \in \mathbf{R}^{k_1 \times n}$ and $\mathbf{A}_y \in \mathbf{R}^{k_2 \times n}$. Note that k_1 and k_2 are the numbers of dictionary atoms for $\mathbf{D}_x$ and $\mathbf{D}_y$, respectively.

In our work, we consider the formulation of sparse representation for E_{DL}, since it has been shown to be very effective in many image synthesis or recognition tasks. For the coupled energy term, we do not explicitly relate the dictionaries $\mathbf{D}_x$ and $\mathbf{D}_y$. Instead, we impose association functions relating the resulting coefficients $\mathbf{A}_x$ and $\mathbf{A}_y$. Once the relationship between $\mathbf{A}_x$ and $\mathbf{A}_y$ is observed, $\mathbf{D}_x$ and $\mathbf{D}_y$ can be updated via E_{DL} accordingly. Therefore, we can convert (1) into the problem below:

$$\begin{aligned}\min_{\mathbf{D}_x, \mathbf{D}_y, \mathbf{A}_x, \mathbf{A}_y} \; & \|\mathbf{X} - \mathbf{D}_x \mathbf{A}_x\|_F^2 + \|\mathbf{Y} - \mathbf{D}_y \mathbf{A}_y\|_F^2 \\ & + \lambda \{\|\mathbf{A}_x\|_1 + \|\mathbf{A}_y\|_1\} + \gamma F(\mathbf{A}_x, \mathbf{A}_y) \\ & \text{s.t. } \|\mathbf{d}_{x,i}\|_2 \leq 1, \|\mathbf{d}_{y,i}\|_2 \leq 1, \forall i,\end{aligned} \tag{2}$$

where λ and γ are the regularization parameters, and $F(\mathbf{A}_x, \mathbf{A}_y)$ is the association function defining the cross-domain relationship in terms of $\mathbf{A}_x$ and $\mathbf{A}_y$. Since our goal is to describe and relate cross-domain data, we now elaborate our determination of $F(\mathbf{A}_x, \mathbf{A}_y)$.

A recent SR work in [25] assumed that LR image patches have the same sparse representations as their HR versions do, and proposed a joint dictionary learning model for representing LR and HR image pairs. Thus, the association function $F(\mathbf{A}_x, \mathbf{A}_y)$ in [25] can be defined as $\|\mathbf{A}_x - \mathbf{A}_y\|_F^2$ with an infinitely large γ. To relax this assumption, Wang *et al.* [19] presented a semi-coupled dictionary learning (SCDL) model and considered $F(\mathbf{A}_x, \mathbf{A}_y) = \|\mathbf{A}_x - \mathbf{W}\mathbf{A}_y\|_F^2$. In other words, SCDL assumes the sparse coefficients from one domain to be identical to those observed at the other domain via a linear projection $\mathbf{W}$.

In order to better describe and associate cross-domain data, we incorporate common feature space learning into the original coupled dictionary learning scheme. In our work, we first replace $F(\mathbf{A}_x, \mathbf{A}_y)$ in (2) by $F(\mathbf{P}_x, \mathbf{P}_y) = \|\mathbf{P}_x - \mathbf{P}_y\|_F^2 = \|\mathbf{U}_x\mathbf{A}_x - \mathbf{U}_y\mathbf{A}_y)\|_F^2$, where $\mathbf{U}_x \in \mathbf{R}^{k_c \times k_1}$ is the projection matrix for $\mathbf{A}_x$, and $\mathbf{P}_x = \mathbf{U}_x\mathbf{A}_x \in \mathbf{R}^{k_c \times n}$ is the projected data of $\mathbf{X}$ in the k_c-dimensional common feature space. The same remarks are applied to $\mathbf{U}_y$ and $\mathbf{P}_y$. It can be seen that we transform the common feature space learning problem into the learning of projection matrices $\mathbf{U}_x$ and $\mathbf{U}_y$, which will be utilized to relate cross-domain data in the derived feature space. Different from prior joint or semi-coupled dictionary learning works, this further relaxes assumptions on the observed dictionaries or sparse coefficients. In other words, instead of minimizing $\|\mathbf{A}_x - \mathbf{A}_y\|_F^2$ or $\|\mathbf{A}_x - \mathbf{W}\mathbf{A}_y\|_F^2$ as [25, 19] did, we consider $F(\mathbf{A}_x, \mathbf{A}_y) = \|\mathbf{U}_x\mathbf{A}_x - \mathbf{U}_y\mathbf{A}_y\|_F^2$ as the association function when solving the coupled dictionary learning problem.

It is worth noting that the solution pair $\mathbf{U}_x$ and $\mathbf{U}_y$ is *not* unique when minimizing $F(\mathbf{P}_x, \mathbf{P}_y) = \|\mathbf{P}_x - \mathbf{P}_y\|_F^2 = \|\mathbf{U}_x\mathbf{A}_x - \mathbf{U}_y\mathbf{A}_y\|_F^2$ (e.g., a trivial solution would be $\mathbf{U}_x = \mathbf{U}_y = \mathbf{0}$). Therefore, we need additional constraints to ensure the uniqueness of $\mathbf{U}_x$ and $\mathbf{U}_y$. In our work, we not only require the common feature space to relate cross-domain data, we also need this space to exhibit additional capabilities in recovering images in one domain using data projected from the other. To be more precise, for an arbitrary instance $\mathbf{p}$ in the common feature space which is projected from the image set $\mathbf{X}$ (or $\mathbf{Y}$), we can derive $\boldsymbol{\alpha}_y = \mathbf{U}_y^{-1}\mathbf{p}$ (or $\boldsymbol{\alpha}_x = \mathbf{U}_x^{-1}\mathbf{p}$) so that the output image in the other domain can be reconstructed by calculating $\mathbf{D}_y\boldsymbol{\alpha}_y$ (or $\mathbf{D}_x\boldsymbol{\alpha}_x$).

From the above observations, we define $F(\mathbf{P}_x, \mathbf{P}_y) = \left\|\mathbf{A}_x - \mathbf{U}_x^{-1}\mathbf{P}_y\right\|_F^2 + \left\|\mathbf{A}_y - \mathbf{U}_y^{-1}\mathbf{P}_x\right\|_F^2$ for the purpose of cross-domain image synthesis. Once the solutions $\mathbf{U}_x$ and $\mathbf{U}_y$ are derived, we have $\mathbf{A}_x \approx \mathbf{U}_x^{-1}\mathbf{P}_y$ and $\mathbf{A}_y \approx \mathbf{U}_y^{-1}\mathbf{P}_x$. It can be seen that, if multiplying both sides by $\mathbf{U}_x$ or $\mathbf{U}_y$, we have $\mathbf{P}_x \approx \mathbf{P}_y$ which implies the minimization of $\|\mathbf{P}_x - \mathbf{P}_y\|_F^2$. This is the reason why the resulting feature space can be considered as a common representation for data from different domains. In our work, we have $k_1 = k_2 = k_c$ since $\mathbf{U}_x$ and $\mathbf{U}_y$ need to satisfy the above function for cross-domain synthesis guarantees. Note that SCDL [19] relates cross-domain data by minimizing $\|\mathbf{A}_x - \mathbf{W}\mathbf{A}_y\|_F^2$, which considers $\mathbf{W}$ as a squared matrix and also has $k_1 = k_2$. The final formulation of our proposed model solves the following optimization problem:

$$\begin{aligned}\min_{\mathbf{D}_x,\mathbf{D}_y,\mathbf{A}_x,\mathbf{A}_y,\mathbf{U}_x,\mathbf{U}_y} & \|\mathbf{X} - \mathbf{D}_x\mathbf{A}_x\|_F^2 + \|\mathbf{Y} - \mathbf{D}_y\mathbf{A}_y\|_F^2 \\ & + \gamma\{\left\|\mathbf{A}_x - \mathbf{U}_x^{-1}\mathbf{P}_y\right\|_F^2 + \left\|\mathbf{A}_y - \mathbf{U}_y^{-1}\mathbf{P}_x\right\|_F^2\} \\ & + \lambda\{\|\mathbf{A}_x\|_1 + \|\mathbf{A}_y\|_1\} + \lambda_R\{\left\|\mathbf{U}_x^{-1}\right\|_F^2 + \left\|\mathbf{U}_y^{-1}\right\|_F^2\} \\ & \text{s.t. } \ \|\mathbf{d}_{x,i}\|_2 \leq 1, \|\mathbf{d}_{y,i}\|_2 \leq 1, \forall i. \end{aligned} \tag{3}$$

In (3), parameters γ and λ balance image representation and sparsity, respectively. We impose additional constraints on $\mathbf{U}_x^{-1}$ and $\mathbf{U}_y^{-1}$ (regularized by λ_R) for numerical stability and to avoid over-fitting.

We would like to point out that, the joint dictionary learning approach in [25] and SCDL in [19] can be viewed as special cases of our proposed model by having $\mathbf{U}_x = \mathbf{U}_y = I$ for [25] or $\mathbf{U}_x = I$ and $\mathbf{U}_y = \mathbf{W}$ for [19]. Nevertheless, our model is more general since we advocate the decomposition/relaxation of $\mathbf{W}$ by learning $\mathbf{U}_x$ and $\mathbf{U}_y$ with bi-directional regularizations. This explains why our model can be applied for solving *both* synthesis and recognition problems. In the next subsection, we will detail the optimization process at the training stage for deriving the dictionary pair, sparse coefficients, and the projection matrices.

2.2. Optimization

While the objective function in (3) is not jointly convex to $\mathbf{D}$, $\mathbf{A}$, and $\mathbf{U}$, it is convex with respect to each of them if the remaining variables are fixed. Given training image data $\mathbf{X}$ and $\mathbf{Y}$, we apply an iterative algorithm (as shown in Algorithm 1) to optimize the dictionaries $\mathbf{D}$, coefficients $\mathbf{A}$, and projection matrices $\mathbf{U}$, respectively. We now discuss how we update these variables in each iteration.

2.2.1 Updating $\mathbf{D}_x$ and $\mathbf{D}_y$

We first apply the approach of joint dictionary learning [25] to calculate $\mathbf{D}_x$ and $\mathbf{D}_y$ for the initialization of the optimization process. When updating the two dictionaries during each iteration, we consider the sparse coefficients $\mathbf{A}$ and projection matrices $\mathbf{U}$ as constants. As a result, the original problem of (3) can be simplified into the following forms:

Algorithm 1 Our Proposed Model

```
Input: Data matrices $\mathbf{X}$ and $\mathbf{Y}$, parameters $\gamma$, $\lambda$, and $\lambda_R$
  1. Initialize $\mathbf{D}^0$ and $\mathbf{A}^0$ by [25], and $\mathbf{U}^0$ as $\mathbf{I}$.
  2. Let $\mathbf{P}_x^0 \leftarrow \mathbf{U}_x^0\mathbf{A}_x^0$ and $\mathbf{P}_y^0 \leftarrow \mathbf{U}_y^0\mathbf{A}_y^0$.
  while not converged do
    3. Update $\mathbf{D}_x^{k+1}$ and $\mathbf{D}_y^{k+1}$ by (4) with $\mathbf{A}_x^k$, $\mathbf{A}_y^k$, $\mathbf{U}_x^k$,
       and $\mathbf{U}_y^k$ derived from the previous iteration.
    4. Update $\mathbf{A}_x^{k+1}$ and $\mathbf{A}_y^{k+1}$ by (5) with $\mathbf{D}_x^{k+1}$, $\mathbf{D}_y^{k+1}$,
       $\mathbf{U}_x^k$, and $\mathbf{U}_y^k$.
    5. Update $\mathbf{U}_x^{k+1}$ and $\mathbf{U}_y^{k+1}$ by (7) with $\mathbf{D}_x^{k+1}$, $\mathbf{D}_y^{k+1}$,
       $\mathbf{A}_x^{k+1}$, and $\mathbf{A}_y^{k+1}$.
    6. $\mathbf{P}_x^{k+1} \leftarrow \mathbf{U}_x^{k+1}\mathbf{A}_x^{k+1}$ and $\mathbf{P}_y^{k+1} \leftarrow \mathbf{U}_y^{k+1}\mathbf{A}_y^{k+1}$
  end while
Output: $\mathbf{D}_x$, $\mathbf{D}_y$, $\mathbf{U}_x$ and $\mathbf{U}_y$
```

Algorithm 2 Cross-Domain Image Synthesis

```
Input: Input $\hat{\mathbf{X}}$; $\mathbf{D}_x$, $\mathbf{D}_y$, $\mathbf{U}_x$ and $\mathbf{U}_y$ trained by Alg. 1.
  1. Initialize $\hat{\mathbf{A}}_x^0$ by (8) and $\hat{\mathbf{A}}_y^0$ by (9).
  2. Let $\hat{\mathbf{P}}_x^0 \leftarrow \mathbf{U}_x\hat{\mathbf{A}}_x^0$, $\hat{\mathbf{P}}_y^0 \leftarrow \mathbf{U}_y\hat{\mathbf{A}}_y^0$, and $\hat{\mathbf{Y}}^0 \leftarrow \mathbf{D}_y\hat{\mathbf{A}}_y^0$
  while not converged do
    3. Update $\hat{\mathbf{A}}_x^{k+1}$ and $\hat{\mathbf{A}}_y^{k+1}$ by (5) with $\hat{\mathbf{Y}}^k$, $\hat{\mathbf{P}}_x^k$, $\hat{\mathbf{P}}_y^k$,
       $\mathbf{U}_x$ and $\mathbf{U}_y$.
    4. Update $\hat{\mathbf{P}}_x^{k+1} \leftarrow \mathbf{U}_x\hat{\mathbf{A}}_x^{k+1}$, $\hat{\mathbf{P}}_y^{k+1} \leftarrow \mathbf{U}_y\hat{\mathbf{A}}_y^{k+1}$,
       and $\hat{\mathbf{Y}}^{k+1} \leftarrow \mathbf{D}_y\hat{\mathbf{A}}_y^{k+1}$
  end while
Output: Output $\hat{\mathbf{Y}}$
```

$$
\begin{aligned}
&\min_{\mathbf{D}_x} \|\mathbf{X} - \mathbf{D}_x\mathbf{A}_x\|_F^2 \quad \text{s.t.} \quad \|\mathbf{d}_{x,i}\|_2 \leq 1, \forall i, \\
&\min_{\mathbf{D}_y} \|\mathbf{Y} - \mathbf{D}_y\mathbf{A}_y\|_F^2 \quad \text{s.t.} \quad \|\mathbf{d}_{y,i}\|_2 \leq 1, \forall i,
\end{aligned} \tag{4}
$$

which is a quadratically constrained quadratic program (QCQP) problem with respect to $\mathbf{D}_x$ or $\mathbf{D}_y$, and the solutions can be solved using Lagrange dual techniques [9].

2.2.2 Updating $\mathbf{A}_x$ and $\mathbf{A}_y$

Similar to dictionary updates, the projection matrices $\mathbf{U}$ and dictionaries $\mathbf{D}$ are fixed when we calculate the solutions of sparse coefficients $\mathbf{A}_x$ and $\mathbf{A}_y$. Besides the standard sparse coding formulation, we have additional terms associated with common feature space learning when updating $\mathbf{A}$. Thus, we convert (3) into the following problem:

$$
\begin{aligned}
&\min_{\mathbf{A}_x} \|\mathbf{X} - \mathbf{D}_x\mathbf{A}_x\|_F^2 + \lambda\|\mathbf{A}_x\|_1 + \gamma\left\|\mathbf{A}_x - \mathbf{U}_x^{-1}\mathbf{P}_y\right\|_F^2, \\
&\min_{\mathbf{A}_y} \|\mathbf{Y} - \mathbf{D}_y\mathbf{A}_y\|_F^2 + \lambda\|\mathbf{A}_y\|_1 + \gamma\left\|\mathbf{U}_y^{-1}\mathbf{P}_x - \mathbf{A}_y\right\|_F^2.
\end{aligned} \tag{5}
$$

To further simplify the above problem, we combine the first and final terms in (5) and rewrite the minimization problem as follows (take $\mathbf{A}_x$ for example):

$$
\min_{\mathbf{A}_x} \left\|\tilde{\mathbf{X}} - \tilde{\mathbf{D}}_x\mathbf{A}_x\right\|_F^2 + \lambda\|\mathbf{A}_x\|_1,
$$

where $\tilde{\mathbf{X}} = \begin{bmatrix} \mathbf{X} \\ \sqrt{\gamma}\,\mathbf{U}_x^{-1}\mathbf{P}_y \end{bmatrix}$ and $\tilde{\mathbf{D}} = \begin{bmatrix} \mathbf{D}_x \\ \sqrt{\gamma}\,I \end{bmatrix}$. This simplified version has the exact formulation as that of the standard sparse coding does. One can simply choose existing solvers like SPAMS [13] for deriving the solutions.

2.2.3 Updating $\mathbf{U}_x$ and $\mathbf{U}_y$

When updating the projection matrices, only the terms associated with $\mathbf{U}_x$ and $\mathbf{U}_y$ in (3) need to be considered into the optimization process. With fixed $\mathbf{D}$ and $\mathbf{A}$, we solve the following ridge regression problems for updating $\mathbf{U}$:

$$
\begin{aligned}
&\min_{\mathbf{U}_x^{-1}} \gamma\left\|\mathbf{A}_x - \mathbf{U}_x^{-1}\mathbf{P}_y\right\|_F^2 + \lambda_R\left\|\mathbf{U}_x^{-1}\right\|_F^2, \\
&\min_{\mathbf{U}_y^{-1}} \gamma\left\|\mathbf{U}_y^{-1}\mathbf{P}_x - \mathbf{A}_y\right\|_F^2 + \lambda_R\left\|\mathbf{U}_y^{-1}\right\|_F^2.
\end{aligned} \tag{6}
$$

From (6), the analytical solutions of $\mathbf{U}$ can be derived as:

$$
\begin{aligned}
\mathbf{U}_x^{-1} &= \mathbf{A}_x\mathbf{P}_y^T(\mathbf{P}_y\mathbf{P}_y^T + (\lambda_R/\gamma)I)^{-1}, \\
\mathbf{U}_y^{-1} &= \mathbf{A}_y\mathbf{P}_x^T(\mathbf{P}_x\mathbf{P}_x^T + (\lambda_R/\gamma)I)^{-1}.
\end{aligned} \tag{7}
$$

To verify that $\mathbf{U}_x^{-1}$ and $\mathbf{U}_y^{-1}$ are invertible, we take $\mathbf{U}_x^{-1}$ for example and need $\mathbf{A}_x\mathbf{P}_y^T = \mathbf{A}_x\mathbf{A}_y^T\mathbf{U}_y^T$ (or $\mathbf{A}_x\mathbf{A}_y^T$) in (7) to be nonsingular. Recall that $\mathbf{A}_x \in \mathbf{R}^{k_1\times n}$ and $\mathbf{A}_y \in \mathbf{R}^{k_2\times n}$ with $k_1 = k_2$. Since we have the number of patches/instances $n \gg k_1$ for image data, it is less likely to have singular $\mathbf{A}_x\mathbf{A}_y^T \in \mathbf{R}^{k_1\times k_1}$. While this has been confirmed by our experiments, one can add small perturbations for inverse guarantees if needed.

Once the optimization is complete, we can apply the derived model for cross-domain image synthesis/recognition.

3. Cross-Domain Image Synthesis & Recognition

We now discuss how we apply the proposed model for solving image synthesis and recognition problems. In particular, examples of single image SR and cross-view action recognition will be presented.

3.1. Cross-domain image synthesis

To address cross-domain image synthesis problems, we first collect cross-domain image/patch pairs for training purposes. Once the training stage is complete, we apply the learned model to synthesize the output image $\hat{\mathbf{Y}}$ from the input image $\hat{\mathbf{X}}$. This is achieved by calculating the sparse coefficients $\hat{\mathbf{A}}_x$ of $\hat{\mathbf{X}}$ via solving

$$
\min_{\hat{\mathbf{A}}_x} \left\|\hat{\mathbf{X}} - \mathbf{D}_x\hat{\mathbf{A}}_x\right\|_F^2 + \lambda\left\|\hat{\mathbf{A}}_x\right\|_1. \tag{8}
$$

Algorithm 3 Cross-Domain Image Recognition

```
Input: Labeled training data $\mathbf{X}$ and unlabeled test data $\mathbf{Y}$.
  $\mathbf{D}$ and $\mathbf{U}$ trained by Alg. 1 using unlabeled data pairs.
  1. Initialize $\mathbf{A}_x^0$ and $\mathbf{A}_y^0$ by (8).
  2. $\mathbf{P}_x^0 \leftarrow \mathbf{U}_x\mathbf{A}_x^0$ and $\mathbf{P}_y^0 \leftarrow \mathbf{U}_y\mathbf{A}_y^0$
  while not converged do
    3. Update $\mathbf{A}_x^{k+1}$ and $\mathbf{A}_y^{k+1}$ by (5) with other variables
       derived from the previous iteration.
    4. $\mathbf{P}_x^{k+1} \leftarrow \mathbf{U}_x\mathbf{A}_x^{k+1}$ and $\mathbf{P}_y^{k+1} \leftarrow \mathbf{U}_y\mathbf{A}_y^{k+1}$
  end while
  5. Train classifiers $\mathbb{C}$ using $\mathbf{P}_x$.
  6. Use $\mathbb{C}$ to predict the labels $\mathbb{L}$ of $\mathbf{P}_y$
Output: $\mathbb{C}$ and $\mathbb{L}$
```

Once $\hat{\mathbf{A}}_x$ is produced, we associate it to $\hat{\mathbf{A}}_y$ by (3) in the derived common feature space:

$$\hat{\mathbf{A}}_y \approx \mathbf{U}_y^{-1}\hat{\mathbf{P}}_x = \mathbf{U}_y^{-1}\mathbf{U}_x\hat{\mathbf{A}}_x. \tag{9}$$

If necessary, one can apply (5) to iteratively update the estimates $\hat{\mathbf{A}}_y$. Finally, we have $\hat{\mathbf{Y}} = \mathbf{D}_y\hat{\mathbf{A}}_y$ as the final synthesized output, as shown in Algorithm 2.

3.2. Cross-domain image recognition

To recognize images at the target domain using labeled source-domain data, we first collect *unlabeled* data pairs from both domains for learning the models $\mathbf{D}$, $\mathbf{A}$, and $\mathbf{U}$. Next, we apply the observed $\mathbf{D}_x$ and $\mathbf{D}_y$ to calculate the sparse coefficients $\mathbf{A}_x$ and $\mathbf{A}_y$ for the labeled source-domain data $\mathbf{X}$ and target-domain test data $\mathbf{Y}$. The matrices $\mathbf{U}_x$ and $\mathbf{U}_y$ then project these coefficients into the common feature space by $\mathbf{P}_x = \mathbf{U}_x\mathbf{A}_x$ and $\mathbf{P}_y = \mathbf{U}_y\mathbf{A}_y$. Finally, classifiers can be designed using $\mathbf{P}_x$ in this feature space, and recognition of $\mathbf{P}_y$ can be performed accordingly. The pseudo code for cross-domain image recognition is shown in Algorithm 3.

As noted in Section 1 and [11], cross-domain recognition approaches based on common feature space learning do *not* necessarily take class label information into their problem formulations (e.g., integrate the stage or regularization term of classifier learning). This is because that, the goal of correspondence-mode approaches like [4, 12] and ours is to derive a common feature space using *only* unlabeled cross-domain data pairs. Once this space is observed, one can project source-domain training (labeled) data and target-domain test data into the derived space, and apply standard classifiers like SVM for recognition.

3.3. Examples

3.3.1 Single-image super resolution

Single-image SR aims at synthesizing a HR image based on one LR input. Although promising SR results have been achieved by example or learning-based methods [5, 25], a major concern is their need to collect training LR and HR image data for designing the SR models. To address this problem, recent approaches like [6, 23] assumed the reoccurrence of patches within and across image scales, so that the SR outputs can be predicted accordingly.

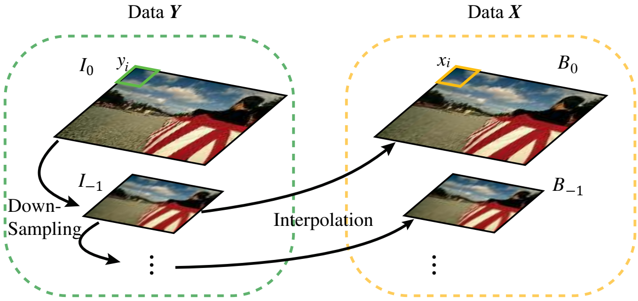

Figure 2. Producing cross-domain data $\mathbf{X}$ and $\mathbf{Y}$ from an input image I_0 (for learning our model for single image super-resolution).

Different from [6, 23], we advance a *self-learning* strategy which constructs cross-domain training data directly from the input image, which allows us to apply our proposed model for solving single-image SR problems. Thus, unlike most learning-based SR approaches, we do not collect training image data beforehand, and no particular post-processing algorithm is required.

Figure 2 shows how we generate cross-domain training data from a LR input I_0. We first construct the image pyramid $\{I_i\}$ by downgrading I_0 into several lower-resolution versions (i.e., I_{-1}, I_{-2}, etc.). With a scaling factor of 2, the size of I_{i-1} is a quarter of that of I_i. In contrast to the pyramid $\{I_i\}$, we upsample the resolution of each I_{i-1} by the same factor to obtain its higher-resolution version B_i. We note that the pyramid $\{I_i\}$ consists of the input image and its downsampled versions, and thus can be considered the ground-truth *target-domain* image set $\mathbf{Y}$. On the other hand, each image B_i is an interpolated version of I_{i-1} (or a blurred version of I_i). Thus, we have $\{B_i\}$ as the *source-domain* image set $\mathbf{X}$. Note that we perform both up/downsampling by bicubic interpolation in our work.

Once image sets $\mathbf{X}$ and $\mathbf{Y}$ are produced, we design our SR model using Algorithm 1. To super-resolve the input LR image I_0, we upsample I_0 into the interpolated version B_1 and consider B_1 as the input image $\hat{\mathbf{X}}$. Finally, Algorithm 2 can be applied to calculate $\hat{\mathbf{Y}}$ for $\hat{\mathbf{X}}$ as the final SR output.

3.3.2 Cross-view action recognition

For cross-view action recognition, one needs to recognize test data captured at one camera using labeled training data at a different view. Recent works like [4, 12, 11] advanced domain adaptation techniques and utilized *unlabeled* data pairs (pre-collected from both camera views) for deriving a common feature space. As a result, training and testing can be performed in this space.

Table 1. Comparisons of PSNR values of different SR approaches.

	airport	airplane	boat	child	lena	man	aerial
bicubic	26.99	25.31	28.19	32.75	27.31	27.12	25.15
ScSR [25]	27.32	26.03	28.72	33.40	27.71	27.77	25.45
SCDL [19]	26.35	24.82	27.9	32.89	27.39	27.04	**26.58**
Glasner [6]	27.28	26.27	28.86	33.48	27.83	27.74	25.57
Ours	**27.76**	**26.79**	**29.63**	**34.29**	**28.51**	**28.42**	26.42

We consider the same setting above and use unlabeled data pairs (e.g., action data *not* of interest) collected by both cameras for learning our model. Once the training is complete, we take labeled source-view data as $\mathbf{X}$ and target-view test data as $\mathbf{Y}$, and we calculate their coefficients $\mathbf{A}_x$ and $\mathbf{A}_y$. Finally, we train classifiers using projected labeled data $\mathbf{P}_x = \mathbf{U}_x\mathbf{A}_x$ in the derived feature space, and perform recognition of $\mathbf{P}_y = \mathbf{U}_y\mathbf{A}_y$ in the same space.

4. Experiments

4.1. Single Image Super-Resolution

We first evaluate the performance of single image SR for cross-domain image synthesis. The images to be super-resolved are collected from the USC-SIPI[1] and Berkeley image segmentation databases [14]. We downgrade the ground-truth HR images with 256 × 256 pixels into 128 × 128 pixels as test LR inputs (as [25] did), and thus the magnification factor is 2 in each dimension. When applying our self-learning scheme to produce cross-domain training data from the LR input, we have the lowest resolution of the image I_i as 32 × 32 pixels (i.e. $i = -2$ in Section 3.3.1). The size of each image patch x_i and y_i in Figure 2 is 5×5 pixels, and the numbers of dictionary atoms for both $\mathbf{D}_x$ and $\mathbf{D}_y$ are $k_1 = k_2 = 512$. We empirically set the regularization parameters $\lambda = \gamma = 0.01$, and $\lambda_R = 0.001$.

We consider the methods of ScSR [25], SCDL [19] and Glasner *et al.* [6] for comparisons. For the method of [6], we apply the code implemented by Yang *et al.* [23]. Since both ScSR and SCDL require training LR and HR image data, we download the code and data from the project websites of [25] and [19]. For fair comparisons, no post-processing is applied to any of the above methods.

Table 1 compares the results of different SR methods in terms of PSNR. It can be seen that our method achieved the highest PSNR values for most of the images, and generally outperformed state-of-the-art SR approaches including ScSR and SCDL. It is worth repeating that, ScSR and SCDL were particularly designed to address image SR, while our model can be applied to both cross-domain synthesis and recognition problems. Thus, our improvements over such methods are appreciable. In addition to PSNR, we also compare the SSIM values of the above approaches. We obtained the highest average SSIM value of 0.8813, while those produced by bicubic, ScSR, SCDL, and Glasner were 0.8526, 0.8675, 0.8562, and 0.8610, respectively. Example SR results are shown in Figures 3∼5 for comparisons.

4.2. Cross-View Action Recognition

We first address cross-view action recognition as one of the cross-domain image recognition tasks. We consider the IXMAS multiview action dataset [21] which contains video frames of eleven action classes. In this dataset, each action video is performed three times by twelve people, and videos of the same action are synchronically captured by five cameras (i.e., cam0 to cam4). Example action videos at different camera views are shown in Figure 6. In our experiments, we choose the same bag-of-features (BOF) model to describe action data as [12] did (the BOF models are calculated from spatial-temporal cuboids extracted from each video at each view using 1000 visual words). Following the same leave-one-action-out strategy as in [12], we take one action class to be recognized, and thus all videos of that action are excluded from the selection of the unlabeled data set. We have $k_1 = k_2 = 50$, and the regularization parameters are also set as $\lambda = \gamma = 0.01$ and $\lambda_R = 0.001$.

Besides CCA which determines a correlation subspace for cross-domain data, we consider two recent approaches of [4, 12] which also focus on deriving common feature spaces for cross-domain recognition. Table 2 compares the performance of different methods, in which the average recognition rates (for all actions) at particular camera-view pairs are listed. For all methods considered, nonlinear SVMs with Gaussian kernels [2] are trained at the derived feature space using labeled data projected from the source view, and recognition is performed on test data projected from the target view. From this table, we see that our approach achieved the highest or comparable recognition results as state-of-the-art methods did.

It is worth repeating that, we consider the setting where only unlabeled cross-domain data pairs are available for learning the domain adaptation model (as [4, 12, 16] did). Therefore, comparisons with methods utilizing label information for associating cross-domain data would be out of the scope of this paper. Nevertheless, the above results confirmed the superiority of our model over CCA and [4, 12].

4.3. Sketch-to-Photo Face Recognition

We now address a more challenging task of sketch-to-photo face recognition, in which features at source and target domains are very different (i.e., sketches vs. photos). In our experiments, a subset of the CUHK Face Sketch Database (CUFS) [20] containing sketch/photo face image pairs of 188 CUHK students is considered (see examples shown in Figure 7). We randomly select 88 sketch-photo pairs as unlabeled data for training our proposed model, and the remaining 100 image pairs are used for evaluating the recognition performance. In particular, the *photo* images of

[1] Available at http://sipi.usc.edu/database.

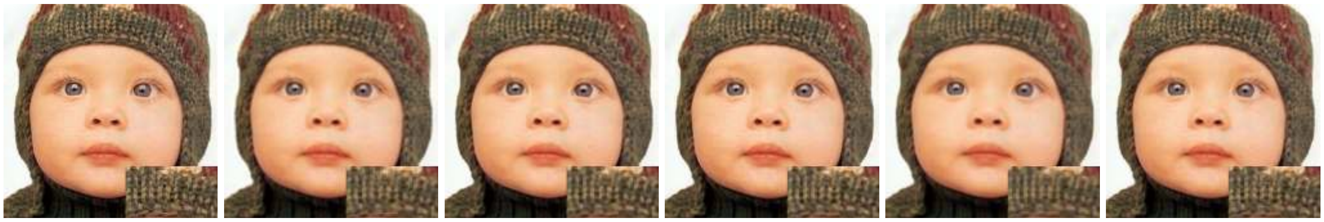

Figure 3. Example SR results and the corresponding PSNR values. Images from left to right: Ground truth, Bicubic (PSNR: 32.75), Glasner *et al.* [6] (PSNR: 33.48), Yang *et al.* [25] (PSNR: 33.40) , Wang *et al.* [19] (PSNR: 32.89) and ours (PSNR: 34.29).

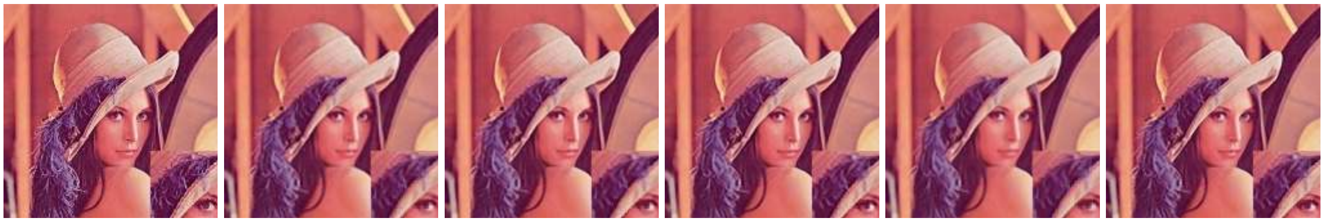

Figure 4. Example SR results and the corresponding PSNR values. Images from left to right: Ground truth, Bicubic (PSNR: 27.31), Glasner *et al.* [6] (PSNR: 27.83), Yang *et al.* [25] (PSNR: 27.45) , Wang *et al.* [19] (PSNR: 27.39) and ours (PSNR: 28.51).

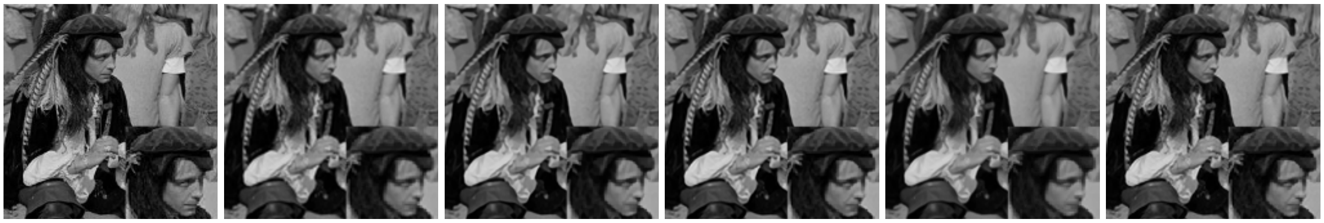

Figure 5. Example SR results and the corresponding PSNR values. Images from left to right: Ground truth, Bicubic (PSNR: 27.12), Glasner *et al.* [6] (PSNR: 27.74), Yang *et al.* [25] (PSNR: 27.77) , Wang *et al.* [19] (PSNR: 27.04) and ours (PSNR: 28.42).

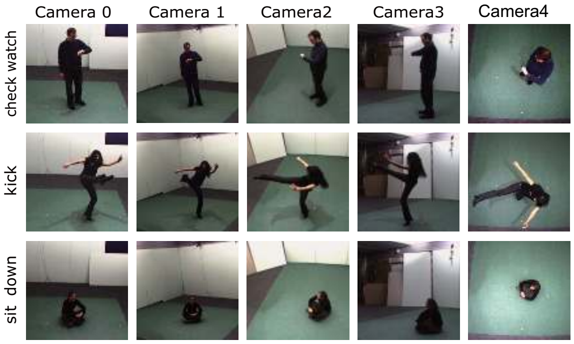

Figure 6. Example actions of the IXMAS dataset. Each row represents an action at five different camera views.

the 100 image pairs are viewed as source domain data and will be projected onto the derived feature space. The corresponding *sketches* will be treated as test data at the target domain for recognition. Once the test images are also projected onto the same feature space, recognition is performed by nearest neighbor (NN) classifiers (as the same classification strategy as [16] did). We repeat the above process five times, and list the average recognition results of different methods in Table 3. We have the same regularization parameters $\lambda = \gamma = 0.01$ and $\lambda_R = 0.001$ for our model.

Besides considering CCA as the baseline approach, we consider the methods of Tang & Wang [17], PLS [16], bilinear model [18], SCDL [19], and joint dictionary learning [25] for comparisons. For SCDL, joint dictionary learning, and our model, we set the numbers of atoms to be learned $k_1 = k_2 = 50$ for the dictionary pair the same at both image domains. For the bilinear model, we select 70 PLS bases and 50 eigenvectors as [16] did. For joint dictionary learning and SCDL, we take the calculated sparse representations as features for performing recognition.

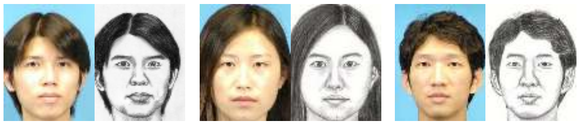

Figure 7. Example sketch-photo image pairs in the CUFS dataset.

From Table 3, it can be seen that our approach achieved the highest recognition performance. It is worth noting that, since the approaches of SCDL and joint dictionary learning were not designed for cross-domain recognition (and did not explicitly derive a common feature space for associating cross-domain data), they are not expected to achieve comparable results as ours does. From the above experiments, the effectiveness of our proposed model for cross-domain image recognition can be successfully verified.

5. Conclusions

We presented a unified model for jointly solving coupled dictionary and common feature space learning prob-

Table 2. Comparisons of recognition rates on the IXMAS dataset. Note that each row corresponds to a source camera view of interest, and each column indicates a target camera view (and the method to be evaluated).

	cam0				cam1				cam2				cam3				cam4			
	CCA	[4]	[12]	Ours	CCA	[4]	[12]	Ours	CCA	[4]	[12]	Ours	CCA	[4]	[12]	Ours	CCA	[4]	[12]	Ours
cam0	–	–	–	–	64.39	72	75.46	**75.76**	66.16	61	64.40	**73.99**	**69.70**	62	67.68	63.89	55.81	30	65.99	**72.48**
cam1	64.90	69	75.72	**76.77**	–	–	–	–	63.89	64	64.23	**68.18**	67.42	68	**68.10**	65.40	54.04	41	56.02	**61.11**
cam2	65.91	62	70.33	**79.04**	61.11	67	66.25	**74.24**	–	–	–	–	66.67	67	71.34	**81.82**	48.99	43	62.42	**66.92**
cam3	65.66	63	**73.74**	71.97	58.08	**72**	65.62	64.90	67.93	68	71.30	**77.78**	–	–	–	–	46.21	44	58.04	**59.85**
cam4	51.01	51	**71.34**	69.44	47.22	55	66.29	**68.94**	54.29	51	**70.88**	69.70	47.98	53	63.55	**65.91**	–	–	–	–

Table 3. Performance comparisons for sketch-to-photo recognition

Tang & Wang [17]	PLS [16]	Bilinear [18]	CCA
81	93.6	94.2	94.6

SCDL [19]	Yang *et al.* [25]	Ours
95.2	95.4	**97.4**

lems. In our work, the derived feature space not only associates cross-domain data for performing recognition, it also updates the dictionaries in each data domain for improved image representation. As a result, the proposed model can be applied to both cross-domain synthesis and recognition problems. From our experiments, we confirmed that our method outperformed state-of-the-art approaches which focused on either learning dictionaries or deriving feature representations for particular cross-domain image synthesis or recognition tasks.

Acknowledgement

This work is supported in part by the Advanced Research Program of ITRI via 102-EC-17-A-01-05-0337, and National Science Council via NSC102-3111-Y-001-015 and NSC102-2221-E-001-005-MY2.

References

[1] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010.

[2] C.-C. Chang and C.-J. Lin. LIBSVM: a library for SVMs. *ACM TIST*, 2001.

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

[4] A. Farhadi and M. Tabrizi. Learning to recognize activities from the wrong view point. In *ECCV*, 2008.

[5] W. T. Freeman, T. Jones, and E. Pasztor. Example-based super-resolution. *IEEE Computer Graphics and Applications*, 2002.

[6] D. Glasner, S. Bagon, and M. Irani. Super-resolution from a single image. In *ICCV*, 2009.

[7] P. Hennings-Yeomans, S. Baker, and B. V. Kumar. Simultaneous super-resolution and recognition of low-resolution faces. In *CVPR*, 2008.

[8] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28:321–377, 1936.

[9] H. Lee, A. Battle, R. Raina, and A. Y. Ng. Efficient sparse coding algorithms. In *NIPS*, 2006.

[10] A. Li, S. Shan, X. Chen, and W. Gao. Maximizing intra-individual correlations for face recognition across pose differences. In *CVPR*, 2009.

[11] R. Li and T. Zickler. Discriminative virtual views for cross-view action recognition. In *CVPR*, 2012.

[12] J. Liu, M. Shah, B. Kuipers, and S. Savarese. Cross-view action recognition via view knowledge transfer. In *CVPR*, 2011.

[13] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *ICML*, 2009.

[14] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*, 2001.

[15] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, 22(10):1345–1359, 2010.

[16] A. Sharma and D. W. Jacobs. Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch. In *CVPR*, 2011.

[17] X. Tang and X. Wang. Face sketch recognition. *IEEE Trans. Circuits Syst. Video Technol.*, 14(1):50–57, 2004.

[18] J. B. Tenenbaum and W. T. Freeman. Separating style and content with bilinear models. *Neural Computation*, 12(6):1247–1283, 2000.

[19] S. Wang, L. Zhang, Y. Liang, and Q. Pan. Semi-coupled dictionary learning with applications in image super-resolution and photo-sketch synthesis. In *CVPR*, 2012.

[20] X. Wang and X. Tang. Face photo-sketch synthesis and recognition. *PAMI*, 31(11):1955–1967, 2009.

[21] D. Weinland, R. Ronfard, and E. Boyer. Free viewpoint action recognition using motion history volumes. *CVIU*, 104(2-3):249–257, 2006.

[22] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *PAMI*, 31(2):210–227, 2009.

[23] C.-Y. Yang, J.-B. Huang, and M.-H. Yang. Exploiting self-similarities for single frame super-resolution. In *ACCV*, 2010.

[24] J. Yang, Z. Wang, Z. Lin, X. Shu, and T. Huang. Bilevel sparse coding for coupled feature spaces. In *CVPR*, 2012.

[25] J. Yang, J. Wright, T. Huang, and Y. Ma. Image super-resolution via sparse representation. *IEEE Trans. Image Process.*, 19(11):2861–2873, 2010.

[26] H. Zhang, J. Yang, Y. Zhang, N. M. Nasrabadi, and T. Huang. Close the loop: Joint blind image restoration and recognition with sparse representation prior. In *ICCV*, 2011.