Subspace Interpolation via Dictionary Learning for Unsupervised Domain Adaptation

Jie Ni, Qiang Qiu and Rama Chellappa
Center for Automation Research
University of Maryland, Collage Park 20742
jni@umiacs.umd.edu, qiu@cs.umd.edu, rama@umiacs.umd.edu

Abstract

Domain adaptation addresses the problem where data instances of a source domain have different distributions from that of a target domain, which occurs frequently in many real life scenarios. This work focuses on unsupervised domain adaptation, where labeled data are only available in the source domain. We propose to interpolate subspaces through dictionary learning to link the source and target domains. These subspaces are able to capture the intrinsic domain shift and form a shared feature representation for cross domain recognition. Further, we introduce a quantitative measure to characterize the shift between two domains, which enables us to select the optimal domain to adapt to the given multiple source domains. We present experiments on face recognition across pose, illumination and blur variations, cross dataset object recognition, and report improved performance over the state of the art.

1. Introduction

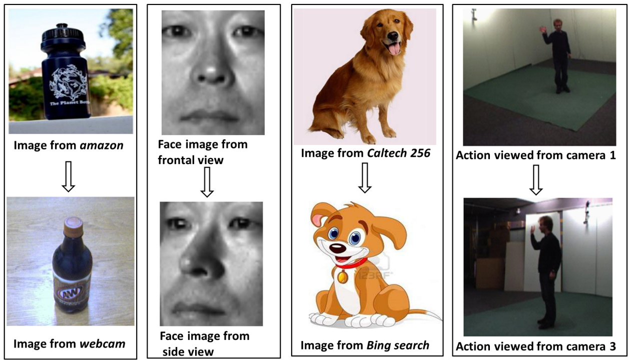
Image from *amazon*
Image from *webcam*
Face image from frontal view
Face image from side view
Image from *Caltech 256*
Image from *Bing search*
Action viewed from camera 1
Action viewed from camera 3

Figure 1. Examples of dataset shifts. Each column contains two images of the same subject collected under different conditions.

Traditional classification problems often assume that training and testing data are captured from the same underlying distribution. Yet this assumption is often violated in many real life applications. For instance, images collected from an internet search engine are compared with those captured from real life [28, 4]. Face recognition systems trained on frontal and high resolution images, are applied to probe images with non-frontal poses and low resolution [6]. Human actions are recognized from an unseen target view using training data taken from source views [21, 20]. We show some examples of dataset shifts in Figure 1.

In these scenarios, magnitudes of variations of innate characteristics, which distinguish one class from another, are oftentimes smaller than the variations caused by distribution shift between training and testing dataset. Directly applying the classifier from the training set to testing set will result in degraded performance. Therefore, it is essential to investigate how to adapt classification systems to new environments. This is often known as the *domain adaptation* problem which has recently drawn much attention in the computer vision community [28, 14, 13, 17].

Domain Adaptation (DA) aims to utilize a *source domain* with plenty of labeled data to learn a classifier for a *target domain* which is collected from a different distribution. Based on the availability of labeled data in the target domain, DA methods can be classified into two categories: *semi-supervised*, and *unsupervised* DA. Semi-supervised DA leverages the few labels in the target data or correspondence between the source and target data to reduce the divergence between two domains. Unsupervised DA is inherently a more challenging problem without any labeled target data to build association between two domains. On the other hand, unsupervised DA is more representative of real-world scenarios. For instance, face recognition systems trained under constrained laboratory environments will encounter great challenges when applied to faces ‘in the wild’, where the acquired face images suffer from a variety of degradations such as low resolution, poor illumination, blur, pose variation, occlusion etc [8]. Sometimes the coupling effects among these different factors give rise to more variations.

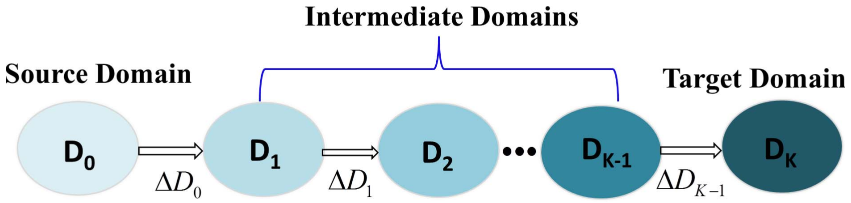

Figure 2. Given labeled data in the source domain and unlabeled data in the target domain, our DA procedure learns a set of intermediate domains (represented by dictionaries $\{\mathbf{D}_k\}_{k=1}^{K-1}$) and the target domain (represented by dictionary $\mathbf{D}_K$) to capture the intrinsic domain shift between two domains. $\{\Delta\mathbf{D}_k\}_{k=0}^{K-1}$ characterize the gradual transition between these subspaces.

As it is very costly to collect labels for target data under various acquisition conditions 'in the wild', it is more desirable that the recognition system be able to adapt in an unsupervised fashion.

An important class of unsupervised DA methods attempts to find suitable representations whose characteristics are shared between the two domains. In this paper, we use subspace representations to model the source and target domains. Subspace modeling has been ubiquitous in the field of computer vision. This is due to the fact that data of high dimensionality usually lie on an intrinsically low-dimensional subspace. In this work, we use a dictionary to represent one domain, as dictionary learning based methods [1, 24] have recently become very popular for subspace modeling. It is based on the fact that data signals in the same subspace can be linearly decomposed with a small number of atoms from an over-complete dictionary. Unlike traditional subspace modeling using Principal Component Analysis (PCA), these atoms are not constrained to be orthogonal, which allows more flexibility to better adapt to the given data signals [23]. The resulting sparse codes are usually leveraged as a feature representation for classification. Effectively learned dictionaries have seen state-of-the-art performance in reconstruction and recognition tasks [11, 32, 22].

Yet the issue of dictionary learning under distribution shifts has received less attention. Specifically, the presence of domain shifts violates the assumption that test data lie in the linear span of training data. As the dictionary atoms learned from one domain are not optimal to fit a different domain, and only a small subset of the atoms are allowed for representation, it will incur large reconstruction errors for the target data. Further, signals of the same class in the target domain will not have similar sparse codes as those from the source domain. These factors will cause inferior performance for both reconstruction and recognition tasks. Therefore, effectively leverage unlabeled target data to adapt the dictionary from one domain to another while maintaining certain invariant representation becomes crucial for successful DA.

We make the following contributions in this paper. (1) We propose a novel unsupervised DA framework by interpolating subspaces through dictionary learning. We hypothesize existence of a virtual path which smoothly connects the source and target domains. Imagine the source domain consists of face images in the frontal view while the target domain contains those in the profile view. Intuitively, face images which gradually transform from the frontal to profile view will form a smooth transition path. Recovering intermediate representations along the transition path allows us to more likely capture the underlying domain shift, as well as to build meaningful feature representations which are preserved across different domains. We encapsulate this intuition into our approach. Specifically, we sample several intermediate domains along a virtual path between the source and target domains, and represent each intermediate domain using a dictionary. We then utilize the good reconstruction property of dictionaries, and learn the set of intermediate domain dictionaries which incrementally reduce the reconstruction residue of the target data. In the mean time, we constrain the magnitude of changes between dictionaries for adjacent intermediate domains to ensure the smoothness of the transition path (refer to Figure 2 for an illustration). (2) We then apply invariant sparse codes across the source, intermediate and target domains to render intermediate representations, which convey a smooth transition in the data signal space. It also provides a shared feature representation where the sample differences caused by distribution shifts are reduced, and we utilize this new feature representation for cross domain recognition. (3) We provide a quantification of domain shift by measuring the similarity between the source and target domain dictionaries which are learned using our DA approach. Presented with multiple domains, this quantitative measure can be exploited to select the optimal domain to adapt to. (4) We demonstrate the wide applicability of our approach for face recognition across pose, illumination and blur variations, cross dataset object recognition, and report the improved performance of our approach over existing DA methods.

Organization of the paper: The structure of the rest of the paper is as follows: In Section 2, we relate our work to existing work on DA. In Section 3, we present our general unsupervised DA approach supported by a quantitative measure of domain shift. We report experimental results on face and object recognition in Section 4. The contributions of the paper are summarized in Section 5.

2. Related work

Several DA methods have been discussed in the literature. We briefly review the relevant work below.

Semi-supervised DA methods rely on labeled target data to perform cross domain classification. Daume [9] proposed a feature augmentation technique such that data points from the same domain are more similar than those from different domains. The Adaptive-SVM method introduced in [33] selects the most effective auxiliary classifiers to adapt to the target dataset. The method in [10] designed a cross-domain classifier based on multiple base kernels. Metric learning approaches [28, 18] were also proposed to learn a cross domain transformation to link two domains. Recently, Jhuo et al. [17] utilized low-rank reconstructions to learn a transformation so that the transformed source samples can be linearly reconstructed by the target samples.

Given no labels in the target domain to learn the similarity measure between data instances across domains, unsupervised DA is more difficult to tackle. Therefore it usually enforces certain prior assumptions to relate source and target data. Structural correspondence learning [7] induces correspondence among features from two domains by modeling their relations with *pivot* features, which appear frequently in both domains. Manifold-alignment based DA [31] computes similarity between data points in different domains through the local geometry of data points within each domain. The techniques in [25, 26] reduce the distance across two domains by learning a latent feature space where domain similarity is measured through maximum mean discrepancy. Shi and Sha [29] define an information-theoretic measure which balances between maximizing domain similarity and minimizing expected classification error on the target domain. Two recent approaches [14], [13] in the computer vision community are more relevant to our methodology, where the source and target domains are linked by sampling finite or infinite number of intermediate subspaces on the Grassmannian manifold. These intermediate subspaces appear to be able to capture the intrinsic domain shift. Compared to their abstract manifold walking strategies, our approach emphasizes on synthesizing intermediate subspaces in a manner which gradually reduces the reconstruction residue of the target data.

Also related is the recent work presented in [27], which jointly learns aligned dictionaries from multiple domains with correspondence available in those domains. Domain invariant sparse codes are designed for cross domain recognition, alignment and synthesis. Our DA approach differs in that we can operate in the unsupervised mode where no correspondence is available.

3. Proposed Method

In this section, we introduce our general framework for unsupervised DA. We first describe some notations to facilitate subsequent discussions.

Let $\mathbf{Y}_s \in \mathbb{R}^{n*N_s}$, $\mathbf{Y}_t \in \mathbb{R}^{n*N_t}$ be the data instances from the source and target domain respectively, where n is the dimension of the data instance, N_s and N_t denote the number of samples in the source and target domains. Let $\mathbf{D}_0 \in \mathbb{R}^{n*m}$ be the dictionary learned from $\mathbf{Y}_s$ using standard dictionary learning methods, e.g, K-SVD [1], where m denotes the number of atoms in the dictionary. As introduced in Section 1, our approach samples several intermediate domains from a smooth transition path between the source and target domains. We associate each intermediate domain with a dictionary $\mathbf{D}_k, k \in [1, K]$, where K is the number of intermediate domains which will be determined in our DA approach.

3.1. Learning Intermediate Domain Dictionaries

Starting from the source domain dictionary $\mathbf{D}_0$, we sequentially learn the intermediate domain dictionaries $\{\mathbf{D}_k\}_{k=1}^K$ to gradually adapt to the target data. This is also conceptually similar to incremental learning. The final dictionary $\mathbf{D}_K$ which best represents the target data in terms of reconstruction error is taken as the target domain dictionary. Given the k-th domain dictionary $\mathbf{D}_k, k \in [0, K-1]$, we learn the next domain dictionary $\mathbf{D}_{k+1}$ based on its coherence with $\mathbf{D}_k$ and the remaining residue of the target data. Specifically, we decompose the target data $\mathbf{Y}_t$ with $\mathbf{D}_k$ and get the reconstruction residue $\mathbf{J}_k$:

$$\mathbf{\Gamma}_k = \arg\min_{\mathbf{\Gamma}} \|\mathbf{Y}_t - \mathbf{D}_k\mathbf{\Gamma}\|_F^2, s.t. \forall i, \|\alpha_i\|_0 \leq T \tag{1}$$

$$\mathbf{J}_k = \|\mathbf{Y}_t - \mathbf{D}_k\mathbf{\Gamma}_k\|_F^2 \tag{2}$$

where $\mathbf{\Gamma}_k = [\alpha_1, ..., \alpha_{N_t}] \in \mathbb{R}^{m*N_t}$ denote the sparse coefficients of $\mathbf{Y}_t$ decomposed with $\mathbf{D}_k$, and T is the sparsity level. We then obtain $\mathbf{D}_{k+1}$ by estimating $\Delta\mathbf{D}_k$, which is the adjustment in the dictionary atoms between $\mathbf{D}_{k+1}$ and $\mathbf{D}_k$:

$$\min_{\Delta\mathbf{D}_k} \|\mathbf{J}_k - \Delta\mathbf{D}_k\mathbf{\Gamma}_k\|_F^2 + \lambda\|\Delta\mathbf{D}_k\|_F^2 \tag{3}$$

Equation (3) consists of two terms. The first term ensures that the adjustments in the atoms of $\mathbf{D}_k$ will further decrease the current reconstruction residue $\mathbf{J}_k$. The second term penalizes abrupt changes between adjacent intermediate domains, so as to obtain a smooth path. The parameter λ controls the balance between these two terms. This is a ridge regression problem. By setting the first order derivatives to be zeros, we obtain the following closed form solution:

$$\Delta\mathbf{D}_k = \mathbf{J}_k\mathbf{\Gamma}_k^T(\lambda\mathbf{I} + \mathbf{\Gamma}_k\mathbf{\Gamma}_k^T)^{-1} \tag{4}$$

where $\mathbf{I}$ is the identity matrix. The next intermediate domain dictionary $\mathbf{D}_{k+1}$ is then obtained as:

$$\mathbf{D}_{k+1} = \mathbf{D}_k + \Delta\mathbf{D}_k \tag{5}$$

Note that when $\lambda = 0$, the Method of Optimal Direction (MOD) [12] becomes a special case of equation (3), where no regularization is enforced.

Starting from the source domain dictionary $\mathbf{D}_0$, we apply the above adaptation framework iteratively, and stop the procedure when the magnitude of $\|\Delta\mathbf{D}_k\|_F$ is below certain threshold, so that the gap between the two domains is absorbed into the learned intermediate domain dictionaries. This stopping criteria also automatically gives the number of intermediate domains to sample from the transition path. We summarize our approach in Algorithm 1. We also show in Proposition 1 that, in each step, the residue $\mathbf{J}_k$ is non-increasing w.r.t the current intermediate domain dictionary and the encoding coefficients. We demonstrate the empirical convergence of our algorithm in Section 4.

Proposition 1. *Given the estimate of $\Delta\mathbf{D}_k$ using equation (4), the residue $\mathbf{J}_k$ is non-increasing w.r.t $\mathbf{D}_k$ and the corresponding sparse coefficients $\mathbf{\Gamma}_k$*

$$\|\mathbf{J}_k - \Delta\mathbf{D}_k\mathbf{\Gamma}_k\|_F^2 \leq \|\mathbf{J}_k\|_F^2 \tag{6}$$

Proof: We provide proof in the appendix.

Algorithm 1 Algorithm to interpolate intermediate subspaces between source and target domains.

```
Input: Dictionary $\mathbf{D}_0$ trained from the source data, target data $\mathbf{Y}_t$, sparsity level $T$, stopping threshold $\delta$, parameter $\lambda$, $k = 0$.
Output: Dictionaries $\{\mathbf{D}_k\}_{k=1}^{K-1}$ for the intermediate domains, dictionary $\mathbf{D}_K$ for the target domain.
while stopping criteria is not reached do
    Decompose the target data with the current intermediate domain dictionary $\mathbf{D}_k$, get the reconstruction residue $\mathbf{J}_k$ using equations (1) and (2)
    Get an estimate of the adjustment in dictionary atoms $\Delta\mathbf{D}_k$ and the next intermediate domain dictionary $\mathbf{D}_{k+1}$ using equations (4) and (5). Normalize the atoms in $\mathbf{D}_{k+1}$ to have unit norm.
    $k \leftarrow k + 1$
    check the stopping criteria $\|\Delta\mathbf{D}_k\|_F \leq \delta$
end while
```

3.2. Recognition Under Domain Shift

Up to now, we have learned a transition path which is encoded with the underlying domain shift. This provides us with rich information to obtain new representations to associate source and target data. Here, we simply apply invariant sparse codes across the source, intermediate, target domain dictionaries $\{\mathbf{D}_k\}_{k=0}^{K}$. The new augmented feature representation is obtained as follows:

$$[(\mathbf{D}_0\alpha)^T, (\mathbf{D}_1\alpha)^T, ..., (\mathbf{D}_K\alpha)^T]^T$$

where $\alpha \in \mathbb{R}^m$ is the sparse code of a source data signal decomposed with $\mathbf{D}_0$, or a target data signal decomposed with $\mathbf{D}_K$. This new representation incorporates the smooth domain transition recovered in the intermediate dictionaries into the signal space. It brings the source and target data into a shared feature space where the data distribution shift is mitigated. Therefore, it can serve as a more robust characteristic across different domains. Given the new feature vectors, we apply PCA for dimension reduction[1], and then employ a SVM classifier for cross domain recognition.

3.3. Quantification of Domain Shift

We now introduce a numeric measure, Quantification of Domain Shift (QDS) to compare the similarity of two domains, which have much practical utility. For instance, we may be faced with more than one source domains in some scenarios. QDS will allow us to select the optimal source domain which has the least domain shift w.r.t the target domain to perform adaptation. We propose to obtain QDS by measuring the similarity between the source domain dictionary $\mathbf{D}_0$ and the target domain dictionary $\mathbf{D}_K$ which is learned using Algorithm 1. This similarity characterizes the amount of domain shift encoded along the transition path. Specifically, it is defined as $Q_{s,t} = \|\mathbf{D}_K^T\mathbf{D}_0\|_F$, where a higher value indicates higher coherence between $\mathbf{D}_0$ and $\mathbf{D}_K$, and less domain shift along the learned transition path. Similarly, by reversing the role of source and target domain to learn the transition path, we can obtain $Q_{t,s}$ which is the amount of shift from target to source domain. Then the symmetric QDS between two domains is defined as $(1/2)(Q_{s,t} + Q_{t,s})$.

4. Experiments

In this section, we evaluate our DA approach on face recognition across pose, lighting and blur variations, and 2D object recognition across different datasets.

4.1. Face Recognition Under Pose Variation

We carried out the first experiment on face recognition across pose variation on the CMU-PIE dataset [30]. We included 68 subjects under 5 different poses in this experiment. Each subject has 21 images at each pose, with variations in lightings. We selected the frontal face images as the source domain, with a total of 1428 images. The

[1] The number of principal components is chosen to preserve 98% of the input data's energy. Alternatively, one can choose any other dimension reduction method for this step.

Table 1. Face recognition under pose variation on CMU-PIE dataset [30]

	c11	c29	c05	c37	average
Ours	76.5	**98.5**	**98.5**	88.2	**90.4**
GFK [13]	63.2	92.7	92.7	76.5	81.3
SGF [14]	58.8	89.7	89.7	72.1	77.6
Eigen light-field [16]	**78.0**	91.0	93.0	**89.0**	87.8
K-SVD [1]	48.5	76.5	80.9	57.4	65.8

target domain contains images at different poses, which are denoted as $c05$ and $c29$ (yawning about $\pm 22.5^o$), $c37$ and $c11$ (yawning bout $\pm 45^o$) respectively. We chose the front-illuminated source images to be the labeled data in the source domain. The task is to determine the identity of the images in the target domain with the same illumination condition. The classification results are in Table 1. We compare our method with the following methods. 1) Baseline K-SVD [1], where target data is directly decomposed with the dictionary learned from the source domain, and the resulting sparse codes are compared using a nearest neighbor classifier. 2) GFK [13] and SGF [14], which perform subspace interpolation via infinite or finite sampling on the Grassmann manifold. 3) Eigen light-field [16] method, which is specifically designed to handle face recognition across pose variations. We observe that the baseline is heavily biased under domain shift, and all DA methods improve upon it. Our method demonstrates its advantage over two other DA methods when the pose variation is large. Furthermore, our average performance is comparable to [16], which relies on a generic training set to build pose specific models, while DA methods do not make such an assumption. We also show some of the synthesized intermediate images in Figure 3 for illustration. As our DA approach gradually updates the dictionary learned from frontal face images using non-frontal images, these transformed representations thus convey the transition process in this scenario. These transformations could also provide additional information for certain applications, e.g. face reconstruction across different poses.

4.2. Face Recognition Across Blur and Illumination Variations

Next, we present the results of a face recognition experiment for dealing with blur and illumination variations. We chose the frontal images of 34 subjects under 21 lighting conditions from the CMU-PIE dataset [30] in this experiment. We selected images of each subject under 11 different illumination conditions to form the source domain. The remaining images with the other 10 illumination conditions were convolved with a blur kernel to form the target domain. Experiments were performed with the Gaussian kernels with standard deviations of 3 and 4, and motion blurs with lengths of 9 (angel $\theta = 135^o$) and 11 (angel $\theta = 45^o$), respectively. We compare our results with those of K-SVD [1], GFK [13] and SGF [14]. Besides, we also compare with the Local Phase Quantization (LPQ) [2] method, which is a blur insensitive descriptor, and the method in [5], which estimates an albedo map (Albedo) as an illumination robust signature for matching. We report the results in Table 2.

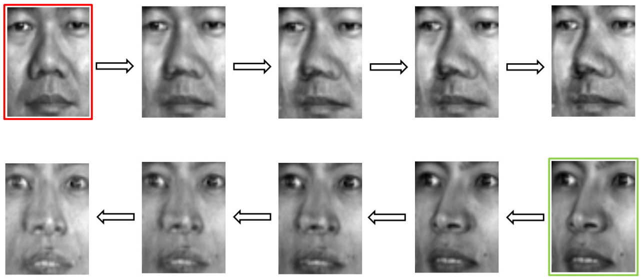

Figure 3. Synthesized intermediate representations between frontal face images and face images at pose $c11$. The first row shows the transformed images from a source image (in red box) to the target domain. The second row shows the transformed images from a target image (in green box) to the source domain.

Table 2. Face recognition across illumination and blur variations on CMU-PIE dataset [30]

	$\sigma = 3$	$\sigma = 4$	$L = 9$	$L = 11$
Ours	**80.29**	**77.94**	**85.88**	**81.18**
GFK [13]	78.53	77.65	82.35	77.65
SGF [14]	70.88	60.29	72.35	67.94
LPQ [2]	66.47	32.94	73.82	62.06
Albedo [5]	50.88	36.76	60.88	45.88
K-SVD [1]	40.29	25.59	42.35	30.59

Our method slightly improves upon GFK [13] and outperforms all other algorithms by a large margin. Since the domain shift in this experiment consists of both illumination and blur variations, traditional methods which are only illumination insensitive or robust to blur are not able to fully handle both variations. DA methods are useful in this scenario as they do not rely on the knowledge of physical domain shift. We also show transformed intermediate representations along the transition path of our approach in Figure 4, which clearly captures the transition from clear to blur images and vice versa. Particularly, we believe that the transformation from blur to clear conditions is useful for

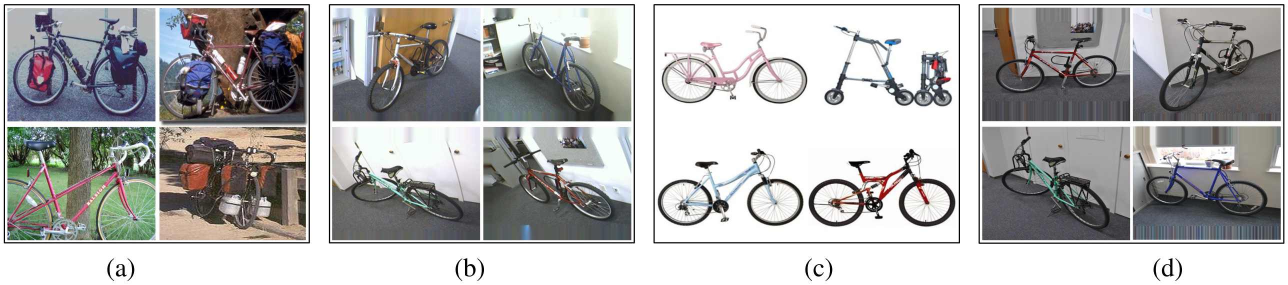

Figure 5. Example images of the *bike* category from the (a) Caltech (b) Webcam (c) Amazon (d) DSLR dataset. (Images best viewed in color)

Table 3. Cross dataset object recognition in both unsupervised and semi-supervised setting

Domain		Unsupervised				Semi-supervised			
source	target	K-SVD [1]	SGF [14]	GFK [13]	Ours	K-SVD [1]	SGF [14]	GFK [13]	Ours
Caltech	Amazon	20.5±0.8	36.8±0.5	40.4±0.7	**45.4**±0.3	31.2±1.0	40.2±0.7	46.1±0.6	**50.0**±0.5
Caltech	DSLR	19.8±1.0	32.6±0.7	41.1±1.3	**42.3**±0.4	34.6±1.0	36.6±0.8	55.0±0.9	**57.1**±0.4
Amazon	Caltech	20.2±0.9	35.3±0.5	37.9±0.4	**40.4**±0.5	25.2±0.7	37.7±0.5	39.6±0.4	**41.5**±0.8
Amazon	webcam	16.9±1.0	31.0±0.7	35.7±0.9	**37.9**±0.9	42.7±0.6	37.9±0.7	56.9 ±1.0	**57.8**±0.5
webcam	Caltech	13.2±0.6	21.7±0.4	29.3±0.4	**36.3**±0.3	23.4±0.4	29.2±0.7	32.8±0.7	**40.6**±0.4
webcam	Amazon	14.2±0.7	27.5±0.5	35.5±0.7	**38.3**±0.3	32.9±0.7	38.2±0.6	46.2±0.7	**51.5**±0.6
DSLR	Amazon	14.3±0.3	32.0±0.4	36.1±0.4	**39.1**±0.5	31.2±1.2	39.2±0.7	46.2±0.6	**50.3**±0.2
DSLR	webcam	46.8±0.8	66.0±0.5	79.1±0.7	**86.2**±1.0	49.9±1.4	69.5±0.9	80.2±0.4	**87.8**±1.0

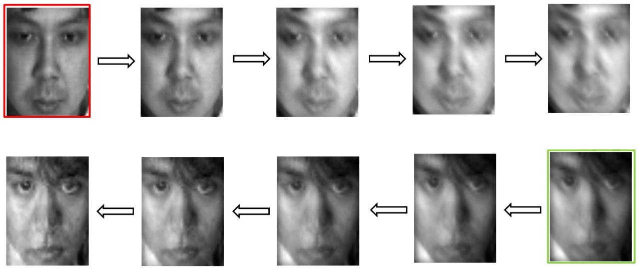

Figure 4. Synthesized intermediate representations from face recognition across blur and illumination variations (motion blur with length of 9). The first row shows the transformed images from a source image (in red box) to the target domain. The second row shows the transformed images from a target image (in green box) to the source domain. (The left most image in the second row is an approximation to the blur-free image in the source domain.)

blind deconvolution, which is a highly under-constrained and costly problem [19].

4.3. Cross Dataset Object Recognition

Following the experiment setting in [13], we evaluated our DA approach for 2D object recognition on four datasets, with a total of 2533 images from 10 categories. The first three datasets were collected by [28], which include images from *amazon.com* (Amazon), collected with a *digital SLR* (DSLR) and a *webcam* (Webcam). The fourth dataset is Caltech-256 (Caltech) [15]. Each dataset constitutes one domain. We used a SURF detector [3] to extract interest points. Then a randomly chosen subset of the interest point descriptors from the Amazon dataset were quantized to visual words by k-means clustering. Each image was represented as a histogram over the quantized visual words of dimension 800. Based on this data representation, we applied our DA approach.

We report performance on eight different pairs of source and target combinations. In the source domain, we randomly selected 8 labeled images per category for Webcam/DSLR/Caltech and 20 for Amazon. Our method is compared with K-SVD [1], GFK [13] and SGF [14]. To draw complete comparison with existing DA methods, we also carried out experiments in the semi-supervised setting where we additionally sampled 3 labeled images per category from the target domain. We ran 20 different trials corresponding to different selections of labeled data from the source and target domains. The average recognition rate and standard deviation was reported in Table 3 for both unsupervised and supervised settings. It is seen that baseline K-SVD has the lowest recognition rate except for one pair of source and target combination in the semi-supervised setting. Overall, our method consistently demonstrates better performance over state-of-the-art methods.

Choice of parameters: In our experiments, the regularization parameter λ varies from 1000 to 2000, and the stopping threshold δ is chosen to be between 0.2 to 0.8.

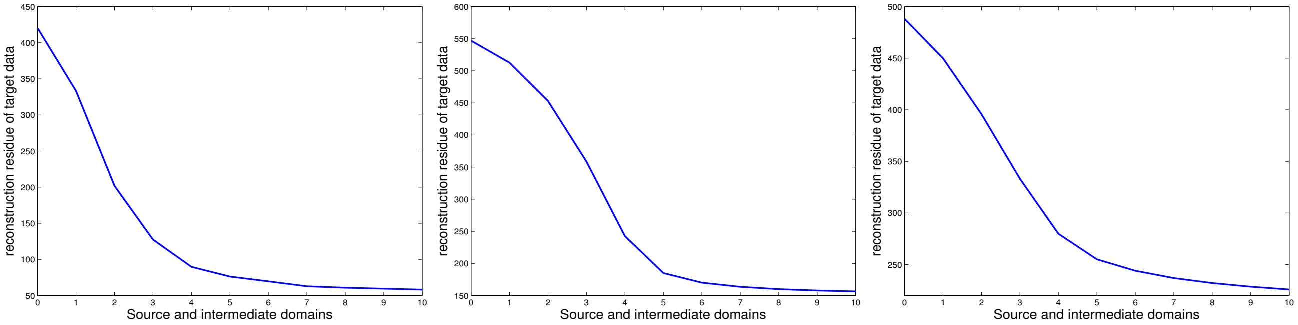

Figure 6. Average reconstruction error of the target domain decomposed with the source and intermediate domains. The combinations of source and target domains are (a) frontal face images v.s. face images at pose $c29$ (b) DSLR v.s. Webcam (c) Caltech v.s. Amazon, respectively.

Table 4. QDS values between Amazon/DSLR/Webcam/Caltech datasets

	Amazon	DSLR	Webcam	Caltech
Amazon	NA	8.13	9.03	**9.78**
DSLR	8.13	NA	**9.60**	8.25
Webcam	9.03	**9.60**	NA	8.96
Caltech	**9.78**	8.25	8.96	NA

Decrease of reconstruction residue along the transition path: Figure 6 shows the average reconstruction residue of target data decomposed with the source, and intermediate domain dictionaries $\{\mathbf{D}_k\}_{k=0}^{K}$ along the transition path which were learned using Algorithm 1. We provide results on three pairs of source and target combinations: frontal face images v.s. face images at pose c29, DSLR v.s. Webcam dataset, Caltech v.s. Amazon, respectively. We observe that the residue is gradually reduced along the transition path, and Algorithm 1 generally stops within five to ten iterations in our experiments, which demonstrates that our framework is able to bridge the gap between two domains.

QDS values: In Table 4, we provide QDS values discussed in Section 3.3 between the Amazon/DSLR/Webcam/Caltech datasets. These quantitative values of domain shift are in line with our experimental performance, i.e., higher QDS values indicate less domain shift, and a higher recognition rate between the corresponding two domains.

5. Conclusions

We presented a fully unsupervised DA method by incrementally learning intermediate domain dictionaries to capture the underlying domain shift. This allows us to transform original data instances from different modalities into a shared feature representation, which serves as a robust signature for cross domain classification. We evaluated our method on public available datasets and obtain improved performance upon the state of the art. We believe our synthesized intermediate representations are also beneficial for certain applications, e.g, face reconstruction across different poses, blur removal etc.

Appendix: Proof of Proposition 1

Substitute (4) into (6), we have

$$\begin{aligned}
&\|\mathbf{J}_k\|_F^2 - \|\mathbf{J}_k - \Delta\mathbf{D}_k\boldsymbol{\Gamma}_k\|_F^2\\
=&\|\mathbf{J}_k\|_F^2 - \|\mathbf{J}_k - \mathbf{J}_k\boldsymbol{\Gamma}_k^T(\lambda\mathbf{I}+\boldsymbol{\Gamma}_k\boldsymbol{\Gamma}_k^T)^{-1}\boldsymbol{\Gamma}_k\|_F^2\\
=&tr(2\boldsymbol{\Gamma}_k^T(\lambda\mathbf{I}+\boldsymbol{\Gamma}_k\boldsymbol{\Gamma}_k^T)^{-1}\boldsymbol{\Gamma}_k\mathbf{J}_k^T\mathbf{J}_k) - \qquad (7)\\
&tr(\boldsymbol{\Gamma}_k^T(\lambda\mathbf{I}+\boldsymbol{\Gamma}_k\boldsymbol{\Gamma}_k^T)^{-1}\boldsymbol{\Gamma}_k\mathbf{J}_k^T\mathbf{J}_k\boldsymbol{\Gamma}_k^T(\lambda\mathbf{I}+\boldsymbol{\Gamma}_k^T\boldsymbol{\Gamma}_k)^{-1}\boldsymbol{\Gamma}_k)
\end{aligned}$$

Let us define the Singular Value Decomposition (SVD) of $\boldsymbol{\Gamma}_k$ as $\boldsymbol{\Gamma}_k = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$, where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices, and $\boldsymbol{\Sigma} = [\tilde{\boldsymbol{\Sigma}}, \mathbf{0}]$ is a rectangular diagonal matrix, with $\tilde{\boldsymbol{\Sigma}} = diag(\sigma_i)$ being a diagonal matrix. Then

$$\begin{aligned}
&\boldsymbol{\Gamma}_k^T(\lambda\mathbf{I}+\boldsymbol{\Gamma}_k\boldsymbol{\Gamma}_k^T)^{-1}\boldsymbol{\Gamma}_k\\
=&\mathbf{V}\boldsymbol{\Sigma}^T\mathbf{U}^T(\lambda\mathbf{I}+\mathbf{U}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\mathbf{U}^T)^{-1}\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T\\
=&[\mathbf{V}_1,\mathbf{V}_2]\boldsymbol{\Sigma}^T\mathbf{U}^T(\lambda\mathbf{I}+\mathbf{U}\tilde{\boldsymbol{\Sigma}}^2\mathbf{U}^T)^{-1}\mathbf{U}\boldsymbol{\Sigma}[\mathbf{V}_1,\mathbf{V}_2]^T \qquad (8)\\
=&\mathbf{V}_1\tilde{\boldsymbol{\Sigma}}(\lambda\mathbf{I}+\tilde{\boldsymbol{\Sigma}}^2)^{-1}\tilde{\boldsymbol{\Sigma}}\mathbf{V}_1^T\\
=&\mathbf{V}_1\boldsymbol{\Phi}\mathbf{V}_1^T
\end{aligned}$$

where $\mathbf{V} = [\mathbf{V}_1, \mathbf{V}_2]$, with $\mathbf{V}_1$ being a square matrix, and $\boldsymbol{\Phi} = diag(\frac{\sigma_i^2}{\sigma_i^2+\lambda})$. Substitute (8) into (7), we have

$$\begin{aligned}
&\|\mathbf{J}_k\|_F^2 - \|\mathbf{J}_k - \Delta\mathbf{D}_k\boldsymbol{\Gamma}_k\|_F^2\\
=&tr(2\mathbf{V}_1\boldsymbol{\Phi}\mathbf{V}_1^T\mathbf{J}_k^T\mathbf{J}_k) - tr(\mathbf{V}_1\boldsymbol{\Phi}\mathbf{V}_1^T\mathbf{J}_k^T\mathbf{J}_k\mathbf{V}_1\boldsymbol{\Phi}\mathbf{V}_1^T)\\
=&tr((2\boldsymbol{\Phi}-\boldsymbol{\Phi}^2)\mathbf{V}_1^T\mathbf{J}_k^T\mathbf{J}_k\mathbf{V}_1) \qquad (9)\\
=&tr(\mathbf{H}\mathbf{V}_1^T\mathbf{J}_k^T\mathbf{J}_k\mathbf{V}_1\mathbf{H})\\
=&\|\mathbf{J}_k\mathbf{V}_1\mathbf{H}\|_F^2 \geq 0
\end{aligned}$$

where $\mathbf{H} = diag(\frac{\sqrt{\sigma_i^4+2\lambda\sigma_i^2}}{\sigma_i^2+\lambda})$

Acknowledgement

We thank the anonymous reviewers for helpful suggestions. This research was supported by a MURI from the US Office of Naval Research under the grant N00014-10-1-0934.

References

[1] M. Aharon, M. Elad, and A. Bruckstein. K-SVD : An algorithm for designing of overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006. 2, 3, 5, 6

[2] T. Ahonen, E. Rahtu, V. Ojansivu, and J. Heikkilä. Recognition of blurred faces using local phase quantization. In *ICPR*, pages 1–4, 2008. 5

[3] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (surf). *Comput. Vis. Image Underst.*, 110(3):346–359, June 2008. 6

[4] A. Bergamo and L. Torresani. Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach. In *NIPS*, pages 181–189, 2010. 1

[5] S. Biswas, G. Aggarwal, and R. Chellappa. Robust estimation of albedo for illumination-invariant matching and shape recovery. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(5):884–899, May 2009. 5

[6] S. Biswas, G. Aggarwal, and P. Flynn. Pose-robust recognition of low-resolution face images. In *CVPR*, pages 601 –608, June 2011. 1

[7] J. Blitzer, R. McDonald, and F. Pereira. Domain adaptation with structural correspondence learning. In *Conference on Empirical Methods in Natural Language Processing*, pages 120–128, 2006. 3

[8] R. Chellappa, J. Ni, and V. M. Patel. Remote identification of faces: Problems, prospects, and progress. *Pattern Recogn. Lett.*, 33(14):1849–1859, Oct. 2012. 1

[9] H. Daumé III. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, June 2007. 3

[10] L. Duan, D. Xu, I. W.-H. Tsang, and J. Luo. Visual event recognition in videos by learning from web data. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(9):1667–1680, 2012. 3

[11] M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 15(12):3736–3745, 2006. 2

[12] K. Engan, S. Aase, and J. Hakon Husoy. Method of optimal directions for frame design. In *ICASSP*, pages 2443–2446, 1999. 4

[13] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, pages 2066–2073, 2012. 1, 3, 5, 6

[14] R. Gopalan, R. Li, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *ICCV*, pages 999–1006, 2011. 1, 3, 5, 6

[15] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical report, Caltech, 2007. 6

[16] R. Gross, I. Matthews, and S. Baker. Appearance-based face recognition and light-fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(4):449–465, April 2004. 5

[17] I.-H. Jhuo, D. Liu, D. T. Lee, and S.-F. Chang. Robust visual domain adaptation with low-rank reconstruction. In *CVPR*, pages 2168–2175, 2012. 1, 3

[18] B. Kulis, K. Saenko, and T. Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *CVPR*, pages 1785–1792, 2011. 3

[19] A. Levin, Y. Weiss, F. Durand, and W. T. Freeman. Understanding and evaluating blind deconvolution algorithms. In *CVPR*, pages 1964–1971, 2009. 6

[20] R. Li and T. Zickler. Discriminative virtual views for cross-view action recognition. In *CVPR*, pages 2855–2862, 2012. 1

[21] J. Liu, M. Shah, B. Kuipers, and S. Savarese. Cross-view action recognition via view knowledge transfer. In *CVPR*, pages 3209–3216, 2011. 1

[22] J. Mairal, F. Bach, and J. Ponce. Task-driven dictionary learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(4):791–804, 2012. 2

[23] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *ICML*, pages 689–696, 2009. 2

[24] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Supervised dictionary learning. In *NIPS*, pages 1033–1040, 2008. 2

[25] S. J. Pan, J. T. Kwok, and Q. Yang. Transfer learning via dimensionality reduction. In *AAAI*, pages 677–682, 2008. 3

[26] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. In *IJCAI*, pages 1187–1192, 2009. 3

[27] Q. Qiu, V. Patel, P. Turage, and R. Chellappa. Domain adaptive dictionary learning. In *ECCV*, pages 631–645, 2012. 3

[28] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *ECCV*, pages 213–226, 2010. 1, 3, 6

[29] Y. Shi and F. Sha. Information-theoretical learning of discriminative clusters for unsupervised domain adaptation. In *ICML*, 2012. 3

[30] T. Sim, S. Baker, and M. Bsat. The cmu pose, illumination, and expression database. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(12):1615–1618, 2003. 4, 5

[31] C. Wang and S. Mahadevan. Manifold alignment without correspondence. In *IJCAI*, pages 1273–1278, 2009. 3

[32] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(2):210–227, Feb 2009. 2

[33] J. Yang, R. Yan, and A. G. Hauptmann. Cross-domain video concept detection using adaptive svms. In *ACM Multimedia*, pages 188–197, 2007. 3