

“© 2016 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.”

# UNSUPERVISED VISUAL DOMAIN ADAPTATION VIA DICTIONARY EVOLUTION

*Songsong Wu<sup>†</sup>, Xiao-Yuan Jing<sup>‡</sup>, Dong Yue<sup>†</sup>, Jian Zhang<sup>‡</sup>, Jian Yang<sup>§</sup>, Jingyu Yang<sup>§</sup>*

<sup>†</sup> School of Automation, Nanjing University of Posts and Telecommunications, China

<sup>‡</sup> School of computer science, Wuhan University, China

<sup>‡</sup> School of Computing and Communications, University of Technology Sydney, Australia

<sup>§</sup> School of Computer Science and Engineering, Nanjing University of Science and Technology, China  
Email: sswuai@gmail.com

## ABSTRACT

In real-world visual applications, distribution mismatch between samples from different domains may significantly degrade classification performance. To improve the generalization capability of classifier across domains, domain adaptation has attracted a lot of interest in computer vision. This work focuses on unsupervised domain adaptation which is still challenging because no labels are available in the target domain. Most of the attention has been dedicated to seeking domain-invariant feature by exploring the shared structure between domains, ignoring the valuable discriminative information contained in the labeled source data. In this paper, we propose a Dictionary Evolution (DE) approach to construct discriminative features robust to domain shift. Specifically, DE aims to adapt a discriminative dictionary learnt based on labeled source samples to unlabeled target samples through a gradual transition process. We show that the learnt dictionary is endowed with cross-domain data representation ability and powerful discriminant capability. Empirical results on real world data sets demonstrate the advantages of the proposed approach over competing methods.

**Index Terms**— Unsupervised domain adaptation, dictionary learning, knowledge transfer

## 1. INTRODUCTION

In machine learning based visual systems, a classifier is learnt from a source domain and applied to a target domain. Recent works [1] show that the classification performance would be dramatically degraded when distribution mismatch between the source and target domains occurs. This problem, known as domain shift, always happens in real world visual applications, e.g. training and test faces in a face recognition task are captured by different lighting conditions and viewing angles. To address domain shift, domain adaptation approaches are

developed to transfer knowledge between domains by making their data distributions consistent. Domain adaptation has become an increasingly important technique in domain shift related visual learning for its promising results in cross-domain object classification [2, 3, 4], object detection [5], video-event detection [6] and face recognition [7].

Domain adaptation assumes that the samples of source and target domains are drawn from a common observation space, while domain shift is caused by the distribution divergence of samples. The main task of domain adaptation is to alleviate the distribution divergence so that the classifier trained based on source samples remains effective for target samples. Two typical approaches of domain adaptation are feature adaptation and classifier adaptation. In feature level, the samples from both domains are represented in a shared feature space where they obey a similar distribution. Then a classifier such as SVM or logistic regression is learnt using labeled samples with expectable generalization performance. The feature space could be obtained by feature transformation [8, 9], feature argumentation [10, 11], or sparse decomposition with a given dictionary [12, 7]. In classification level, domain adaptation is achieved by designing an objective function for typical classifiers such as SVM [6], multiple kernel learning [13] to reduce the distribution mismatch. A recent work [14] tries to tackle domain shift by jointly learning domain-invariant feature and classification model.

Domain adaptation also assumes that there are plenty of labeled samples in source domain, and few or no labeled samples in the target domain. According to the availability of labeled samples in the target domain, domain adaptation approaches fall into two categories: semi-supervised domain adaptation and unsupervised domain adaptation. Semi-supervised domain adaptation usually constructs learning model using the priori information from the class labels of the source and target samples [11], or the pairwise similarity between them [3, 4]. As no class label nor similarity information of the source and target samples is available, unsupervised domain adaptation often exploits the underlying geometry structure [15, 16] or the intrinsic probability distribution

---

This work is supported by Natural Science Foundation of China: 61402238 and 61272273, Jiangsu Postdoctoral Sustentation Fund, China: 1302054C, and Research Project of NJUPT: XJKY14010 and XJKY14016.

of data across domains to bridge the gap of domains [9, 17]. In this paper, we focus on the issue of unsupervised visual domain adaptation, which is more common in real world applications and relatively more challenging as well.

As the association between domains is hard to build without any labeled target samples, how to learn sample representation capturing the shared characteristics of domains is crucial to unsupervised domain adaptation. In this context, dictionary learning and sparse representation approaches have attract a lot of interest. Recent studies include [12], which proposed to interpolate a serial of dictionaries to characterize the intermediate domains between the source and target, and [18], which proposed to learn a shared dictionary to reduce the mismatch of data distribution and preserve geometry structure of data. Although these new methods have shown to be better than the state of the art, there remains a common issue that limit their performances: the discriminative ability of the desired dictionary are largely ignored. So it is desirable to explore the discriminant information from the source domain to facilitate domain adaptation. We claim that a shared dictionary joining discriminative power with representative ability is more conducive to cross-domain visual recognition.

We propose a novel Dictionary Evolution (DE) approach in this paper to achieve cross domain discriminant information transfer. Specifically, our contributions include: (1) we develop a gradual dictionary iteration process in which a dictionary specific to the source domain evolves into a shared dictionary with both representation and discriminative power; (2) based on the shared dictionary, we develop a domain-invariant discriminative feature extraction scheme; (3) Through extensive experiments, we demonstrate our DE approach can be widely applied to cross domain object classification, cross dataset digit recognition, face recognition across pose, and report the improved performance of DE over existing methods for unsupervised visual domain adaptation.

## 2. CROSS-DOMAIN DICTIONARY EVOLUTION

### 2.1. Problem Statement

According to [19], a domain  $\mathcal{D}$  is defined by a feature space  $\mathcal{Y}$  and a marginal probability distribution  $P(\cdot)$ , i.e.  $D = \{\mathcal{Y}, P(y)\}$ , where  $y \in \mathcal{Y}$ . A task  $\mathcal{T}$  associated with a specific domain  $\mathcal{D}$  is defined by a label space  $\mathcal{L}$  and a prediction function  $f(\cdot)$ , i.e.  $\mathcal{T} = \{\mathcal{L}, f(y)\}$ , where  $f: \mathcal{Y} \mapsto \mathcal{L}$ . In the context above, traditional learning problem is stated as the estimation of function  $f(y)$  merely using training sample pairs  $\{(y_i, l_i)\} \in \mathcal{Y} \times \mathcal{L}$ .

Domain adaptation considers two domains  $\mathcal{D}_s$  (referred as source) and  $\mathcal{D}_t$  (referred as target) with the relations:  $\mathcal{Y}_s = \mathcal{Y}_t$ ,  $\mathcal{L}_s = \mathcal{L}_t$ ,  $P_s(y_s) \neq P_t(y_t)$ , and  $f_s(y_s) \neq f_t(y_t)$ . The goal is to estimate the target predictive function  $f_t(\cdot)$  for target samples  $\{y_{t_i}\}_{i=1}^{nt}$  with the aid of source sample pairs  $\{(y_{s_i}, l_{s_i})\}_{i=1}^{ns}$ .

In this paper, we try to find a domain-invariant feature space in which distribution matching and classifier compatibility between domains are achieved. Specifically, our goal is to learn a shared dictionary for domains leading to a discriminative representation of samples satisfying  $P_s(\cdot) = P_t(\cdot)$  and  $f_s(\cdot) \approx f_t(\cdot)$ .

### 2.2. The Learning Model

Our main idea is to adapt increasingly a discriminative dictionary  $D_0$ , learnt based on source samples, to target samples through a dictionary evolution process. During the evolution process, the dictionary is required to enhance the power of data representation for samples and retain discriminative capability. Let  $Y_s$  and  $Y_t$  be source and target sample sets respectively. At the  $k$ -step of dictionary evolution process, dictionary  $D_k$  is required to provide sparse coding  $X_s$  and  $X_t$  for source and target domains with low reconstruction error and strong discriminative capability. To this end, we propose the following learning model

$$\arg \min_{(D_k, X_s, X_t)} F_t(D_k, X_t) + \lambda F_s(D_k, X_s) \quad (1)$$

The first term controls the effectiveness of  $D_k$  to represent target samples by sparse coding. The second term is the loss of discriminative information during evolution process.  $\lambda$  control the tradeoff between cross-domain representation and discriminative capability of the learnt dictionary. Below, we discuss the design of objective functions  $F_t(D_k, x_t)$  and  $F_s(D_k, X_s)$  used in each evolution step. To facilitate the narrative, we abbreviate  $F_t(D_k, X_t)$  as  $F_t^k$ ,  $F_s(D_k, X_s)$  as  $F_s^k$ , and define  $F^k = F_t^k + \lambda F_s^k$  throughout the rest of the paper.

### 2.3. Target Domain Oriented Objective

The initial dictionary  $D_0$  is learnt to represent source data, while it can barely represent target data due to distribution divergence between domains. To address this issue, we propose to amend the dictionary gradually by  $D_{new} = D_{old} + \Delta D$  so that lower reconstruction error to  $Y_t$  could be obtained by  $D_{new}$  and the corresponding sparse coding. In particular, at the  $k$ th step of dictionary evolution, we aim to search the change in dictionary  $\Delta D_k$  by minimizing

$$F_t^k = \|Y_t - (D_{k-1} + \Delta D_k)X_t^k\|_F^2 + \alpha_t \sum_{i=1}^{nt} \|x_{t_i}^k\|_1 \quad (2)$$

where  $X_t^k = [x_{t_1}^k, \dots, x_{t_{nt}}^k]$  is the sparse coding of target samples based on dictionary  $D_k$ , and  $\alpha_t$  controls the sparsity degree of  $x_{t_i}$ .

### 2.4. Source Domain Oriented Objective

For multiclass discrimination, the dictionary in evolution should meet two conditions. First, it's able to provide sparse

representation to optimally reconstruct the source data so that the dictionary is robust to domain shift. Second, it can produce similar sparse representations of source data as those produced by  $D_0$  so that the original discriminative information can be preserved. To this end, we propose the objective function  $F_s$  as

$$F_s^k = \left\{ \begin{array}{l} \|Y_s - (D_{k-1} + \Delta D_k)X_s^k\|_F^2 + \\ \alpha_s \sum_{i=1}^{n_s} \|x_{s_i}^k\|_1 + \beta \|X_s^k - X_s^0\|_F^2 \end{array} \right\} \quad (3)$$

where  $X_s^k = [x_{s_1}^k, \dots, x_{s_{n_s}}^k]$  is the sparse coding of source samples based on dictionary  $D_k$ , and  $\alpha_s$  controls the sparsity degree of  $x_{s_i}$ , and  $\beta$  controls the weight of discriminative information preservation.

The basic principle of the last term in (3) is that the similarity between dictionaries can be transformed as the similarity between the corresponding sparse representations. Thus, the last term means that the sparse representation based on  $D_k$  is discriminative, which indicates the discriminativeness of the atoms in  $D_k$ .

## 2.5. Optimization Algorithm

By incorporating Eqs.(2) and (3) into Eq.(1), we have the following DDE model:

$$\arg \min_{(\Delta D_k, X_s^k, X_t^k)} \left\{ \begin{array}{l} \|R_t^{k-1} - \Delta D_k X_t^k\|_F^2 + \alpha_t \|X_t^k\|_1 + \\ \lambda (\|R_t^{k-1} - \Delta D_k X_s^k\|_F^2 + \\ \alpha_s \|X_s^k\|_1 + \beta \|X_s^k - X_s^0\|_F^2) \end{array} \right\} \quad (4)$$

where  $R_t^{k-1} = Y_t - D_{k-1}X_t^k$  and  $R_s^{k-1} = Y_s - D_{k-1}X_s^k$  are reconstruction residues based on  $D_{k-1}$ ;  $\|X_t^k\|_1 = \sum_{i=1}^{n_t} \|x_{t_i}^k\|_1$  and  $\|X_s^k\|_1 = \sum_{i=1}^{n_s} \|x_{s_i}^k\|_1$  denote the sum of  $l_1$ -norm of sparse codings.

The objective function in Eq.(4) is not convex for  $D_k$ ,  $X_s^k$  and  $X_t^k$  simultaneously, but fortunately it is convex to one of them when the other two are fixed. We search the optimal  $D_k$ ,  $X_s^k$  and  $X_t^k$  iteratively by using the following three-stage update process.

**Learning  $D_k$  with fixed  $X_s^k$  and  $X_t^k$**  By fixing  $X_s^k$  and  $X_t^k$ , Eq.(4) can be rewritten as

$$\arg \min_{\Delta D_k} \|R_t^{k-1} - \Delta D_k X_t^k\|_F^2 + \lambda \|R_t^{k-1} - \Delta D_k X_s^k\|_F^2 \quad (5)$$

Eq.(5) is a least square problem whose optimal solution is given by

$$\Delta D_k^* = (R_t^{k-1}(X_t^k)^T + \lambda R_s^{k-1}(X_s^k)^T)(X_t^k(X_t^k)^T + \lambda X_s^k(X_s^k)^T)^{-1} \quad (6)$$

**Learning  $X_t^k$  with fixed  $D_k$  and  $X_s^k$**  By fixing  $D_k$  and  $X_s^k$ , Eq.(4) can be rewritten as

$$\arg \min_{X_t^k} \|R_t^{k-1} - \Delta D_k X_t^k\|_F^2 + \alpha_t \|X_t^k\|_1 \quad (7)$$

Eq.(7) is a typical sparse coding problem. Here we follow the work of [20] to obtain the optimal  $X_t^k$  by using a variant version of the Iterative Projection Method (IPM). The derivative of the objective term  $\|R_t^{k-1} - \Delta D_k X_t^k\|_F^2$  w.r.t  $X_t^k$  is given by

$$\nabla_{X_t^k} = 2(D_k^T D_k X_t^k - D_k^T Y_t^k) \quad (8)$$

**Learning  $X_s^k$  with fixed  $D_k$  and  $X_t^k$**  When  $D_k$  and  $X_t^k$  are fixed, Eq.(4) can be rewritten as

$$\arg \min_{X_s^k} \|R_s^{k-1} - \Delta D_k X_s^k\|_F^2 + \alpha_s \|X_s^k\|_1 + \beta \|X_s^k - X_s^0\|_F^2 \quad (9)$$

Eq.(9) is also a sparse coding problem that can be solved by the IPM. The derivative of the objective term  $\|R_s^{k-1} - \Delta D_k X_s^k\|_F^2 + \beta \|X_s^k - X_s^0\|_F^2$  w.r.t  $X_s^k$  is given by

$$\nabla_{X_s^k} = 2(D_k^T D_k X_s^k - D_k^T Y_s^k + \beta(X_s^k - X_s^0)) \quad (10)$$

The above three alternative optimizations are convex, so we can obtain the optimal solutions of Eq.(4) until the iteration is convergent. The algorithm of dictionary evolution is summarized in Algorithm 1.

---

### Algorithm 1 Dictionary Evolution

---

**Require:** source samples  $X_s$ , target samples  $X_t$ , initial dictionary  $D_0$ , tradeoff parameter  $\lambda$ , sparse coding parameter  $\alpha_s$  and  $\alpha_t$ , regularization parameter  $\beta$ .

**Ensure:** dictionary  $D_k$ , sparse representations  $X_s^k$ , and  $X_t^k$

$X_s^0 \leftarrow$  by sparse decomposition for  $X_s$  with  $D_0$

$X_t^0 \leftarrow$  by sparse decomposition for  $X_t$  with  $D_0$

**for all**  $k \geq 1$  such that  $F^k \leq F^{k-1}$  **do**

$X_s^k \leftarrow X_s^{k-1}$ ,  $X_t^k \leftarrow X_t^{k-1}$

**while not convergent do**

$D_k \leftarrow$  by Eq.(6);

$X_t^k \leftarrow$  by solving sparse coding problem Eq.(7);

$X_s^k \leftarrow$  by solving sparse coding problem Eq.(9);

**end while;**

**end for;**

**return**  $D_k$ ,  $X_s^k$ , and  $X_t^k$ .

---

## 2.6. Classification Scheme

To use  $D_k$  for visual classification, we propose the following scheme including four steps:

**Step 1** Reconstruct the source and target samples by  $\tilde{Y}_s = D_k X_s^k$  and  $\tilde{Y}_t = D_k X_t^k$  respectively, then form a training set  $\tilde{Y}_{train} = [\tilde{Y}_s, \tilde{Y}_t]$

**Step 2** Perform PCA on  $\tilde{Y}_{train}$  to learn a linear projection  $P$  by preserving %98 variance energy of training samples<sup>1</sup>.

<sup>1</sup>Any other dimension reduction method can be chosen for this step.

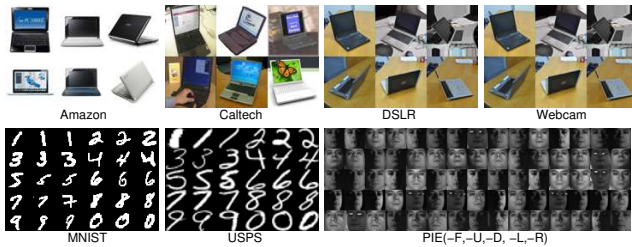
**Step 3** For a test target sample,  $y_{test}$ , compute its sparse representation with  $D_K$  then reconstruct it as  $\tilde{y}_{test}$ . Project  $\tilde{y}_{test}$  onto the subspace spanning by  $P$  to get the corresponding feature.

**Step 4** A classifier trained from the source features is applied to classify the test sample<sup>2</sup>.

### 3. EXPERIMENTAL RESULTS

#### 3.1. Datasets and Experimental settings

We conduct experiments on Office+Caltech, USPS+MNIST and PIE(refer to Figure1) which are benchmark datasets widely adopted to evaluate visual domain adaptation algorithms.



**Fig. 1.** Examples of Office+Caltech, USPS+MNIST and PIE

**Office+Caltech** We adopt the public Office+Caltech datasets released by [2] consisting of 2533 images from the 10 object classes common to all four datasets, in which the images is represented by SURF descriptors quantized into histograms of 800 bag-of-visual words and standardized by z-score normalization. For the four domains, C (Caltech-256), A (Amazon), W (Webcam), and D (DSLR), all the 12 possible source-target domain pairs are considered. We use the data splits suggested by [12].

**USPS+MNIST** USPS [21] consists of 7,291 training images and 2,007 test images of size  $16 \times 16$ . MNIST [22] has a training set of 60,000 examples and a test set of 10,000 examples of size  $28 \times 28$ . They share 10 classes of digits. To speed up experiments, we construct one dataset USPS vs MNIST by randomly sampling 1,800 images in USPS to form the source data, and randomly sampling 2,000 images in MNIST to form the target data. We switch source/target pair to get another dataset MNIST vs USPS. We uniformly rescale all images to size  $16 \times 16$ , and represent each one by a feature vector encoding the gray-scale pixel values.

**PIE** CMU-PIE dataset [23] has 68 individuals with face images captured by different poses and different illuminations and/or expressions. In these experiments, we adopt five subsets of PIE, each corresponding to a different pose. Specifically, we choose PIE-F (C27, frontal pose) as the source domain, PIE-L (C05, left pose), PIE-U (C07, upward pose), PIE-D

(C09, downward pose) and PIE-R (C29, right pose) as the target domains. In each domain (pose), 15 images of each individual are randomly sampled. In this way, the source and target data are constructed using face images from different poses, thus will follow significantly different distributions. To shorten experiment time, the images are resized to  $16 \times 16$ .

#### 3.2. Comparison Methods

The proposed DE approach is compared with two baseline methods and four state-of-the-art (related) methods for unsupervised domain adaptation: Fisher Discriminative Dictionary Learning (**FDDL**) [20], Geodesic Flow Kernel (**GFK**) [2], Subspace Alignment (**SA**) [24], Subspace Interpolation via Dictionary Learning (**SIDL**) [12], Transfer Sparse Coding (**TSC**) [18].

In particular, SIDL and TSC are two dictionary learning based feature adaptation method that aim to achieve robustness of sparse coding with a shared dictionary. In our experiments, the initial dictionary of SIDL is provided by performing K-SVD(referred as SIDL1) or FDDL(referred as SIDL2) in the source domain.

For classification, a SVM with Gaussian kernel is used with all the methods but FDDL, which is developed with a specific classification scheme [20]. The width parameter of Gaussian kernel is determined by method [25], and the C parameter of SVM is tuned by 5-fold cross-validation on the source over the range  $\{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$ . Our DE method has four main parameters ( $\lambda, \alpha_s, \alpha_t, \beta$ ) that are also chosen by 5-fold cross validation on the source. Note that as the target data are unlabeled, it is only feasible to tune the parameters on the source. In particular, the parameter  $\lambda$  was tuned in  $\{0.1, 0.2, \dots, 1\}$ ,  $\beta$  in  $\{10, 20, \dots, 90, 100, 200, \dots, 1000\}$ . The sparse parameter  $\alpha_s$  and  $\alpha_t$  are tuned in 0.05 and 0.1 respectively according to the offered initial dictionary  $D_0$ .

#### 3.3. Experimental Results

Table1 shows a comparison of the recognition accuracies of different methods on 18 source-target domain pairs. The proposed DE algorithm outperforms the considered methods in 13 pairs out of 18 and gains performance improvements in average accuracy of 1.64%, 1.19%, and 1.79% compared to the best method for comparison on the three datasets respectively. The results demonstrate that DE is effective to adapt a discriminative dictionary from the source domain to the target domain, thereby producing domain-invariant feature with discriminative power.

We observe that DE significantly outperforms TSC and SIDL, which are two state-of-the-art sparse coding based unsupervised domain adaptation methods. Instead of only laying emphasis on the representation ability of dictionary as TSC and SIDL do, DE pay additional attention to discriminatory information transfer between different domains, which

<sup>2</sup>The feature is expected to be compact, discriminative and robust to domain shift, which benefits from the shared dictionary.

**Table 1.** Accuracy(%) on cross-domain visual datasets(bold numbers indicate the best results).

Dataset	FDDL	GFK	SA	SIDL1	SIDL2	TSC	DE
A→C	33.10	40.49	39.49	40.20	39.78	36.79	<b>40.94</b>
A→D	32.90	35.10	35.61	<b>37.23</b>	36.78	31.75	36.62
A→W	32.14	<b>39.73</b>	38.20	38.64	38.98	37.20	38.66
C→A	33.03	40.92	38.27	37.49	38.88	40.26	<b>41.11</b>
C→D	33.85	38.95	37.99	38.34	37.90	34.14	<b>39.49</b>
C→W	28.78	33.92	31.49	34.29	33.22	30.53	<b>36.22</b>
D→A	31.47	37.20	34.71	39.03	37.68	35.94	<b>40.66</b>
D→C	29.46	33.34	34.46	34.16	33.54	33.68	<b>34.38</b>
D→W	74.37	74.17	68.00	77.39	73.64	<b>79.54</b>	79.05
W→A	32.24	36.26	36.30	34.82	37.16	37.28	<b>37.41</b>
W→C	27.56	30.53	31.06	29.02	30.83	<b>32.63</b>	31.21
W→D	69.39	69.11	57.32	69.55	66.53	71.21	<b>74.11</b>
Average	38.19	42.48	40.24	42.51	42.08	41.75	<b>44.16</b>
M→U	59.39	23.89	51.44	53.72	52.22	61.56	<b>62.28</b>
U→M	47.00	31.25	46.40	39.65	32.55	52.50	<b>54.15</b>
Average	53.20	27.57	48.92	46.69	42.39	57.03	<b>58.22</b>
P:F→U	69.41	64.71	66.08	68.53	66.57	68.33	<b>71.18</b>
P:F→D	69.12	70.59	71.57	<b>74.80</b>	69.22	70.88	72.45
P:F→L	50.20	43.04	45.98	49.12	47.94	53.43	<b>55.00</b>
P:F→R	38.53	40.29	42.25	42.45	41.18	42.75	<b>43.92</b>
Average	56.82	54.66	56.47	58.73	56.23	58.85	<b>60.64</b>

is more favorable to visual recognition. We also notice that SIDL2 underperforms SIDL1 in general, outperforms FDDL just on 1 out of 3 datasets. This indicates that the dictionary interpolation principle of SIDL helps to transfer the representation capability of a dictionary to new domains, but can not guarantee the transfer of discriminative ability of dictionary. This meanwhile indirectly verifies the effectiveness of the discriminant analysis in our dictionary learning model.

### 3.4. Empirical analysis of algorithm property

We conducted experiments for parameter sensitivity and convergence property analysis of the DE algorithm. Figure 2 shows the results on  $W \rightarrow D$ ,  $U \rightarrow M$  and  $P:F \rightarrow L$ , while similar trends on all other datasets are omitted due to space limitation. The implications are briefly discussed below

1. **Parameter  $\lambda$ :** Theoretically,  $\lambda$  controls the weight of discriminative item in the objective of DE. Larger values of  $\lambda$  will produce obstructs to dictionary evolution causing the lack of cross-domain robust, while smaller values of  $\lambda$  may cause trouble for matrix inverting in formular(6). We plot the classification accuracy w.r.t. different values of  $\lambda$  in Figure2(a), and can choose  $\lambda \in \{0.1, 0.2, \dots, 1\}$ .
2. **Parameter  $\beta$ :** Intuitively, smaller values of  $\beta$  will lead to deficiency of discriminative power of dictionary atoms, while with larger values of  $\beta$ , DE tends to left the evolved dictionary keep the same as the initialization dictionary. We plot classification accuracy w.r.t. different values of  $\beta$  in Figure2(b), and choose  $\beta \in \{10, 20, \dots, 100, 200, \dots, 1000\}$ .

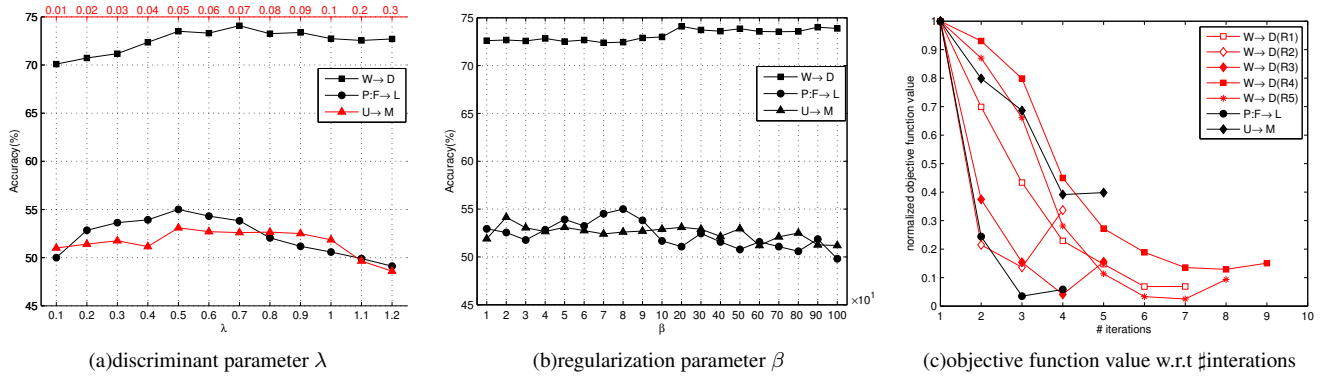
3. **Convergence:** We also empirically check the convergence property of DE. Figure2(c) shows that the objective function value associated with the intermediate dictionary during evolution decreases steadily. We observe that dictionary evolution can reach the balance between cross-domain representation and discriminaiveness shortly after several iterations. After that the objective function value increases, which terminates dictionary evolution due to the terminating condition of DE, so the optimal dictionary being in equilibrium is obtained. Note that because there are differences in magnitude of objective function value on the three datasets, we normalize them according to the maximum objective function value on each dataset.

## 4. CONCLUSION

Pervious unsupervised domain adaptation methods based on sparse coding and dictionary learning exclusively focus on the cross-domain representation ability of dictionary. We argue that the discriminativeness of sparse representation is important to cross-domain feature adaptation, and present a novel Dictionary Evolution (DE) approach to learn a shared dictionary between different domains with discriminative capability. Specifically, the evolution direction of DE is controlled by both the cross-domain representational ability and discriminative power. Extensive experimental results on the object, digit and face recognition databases have shown that, by taking advantage of dictionary learning with discrimination information transfer, DE outperforms other competing methods for unsupervised visual domain adaptation.

## 5. REFERENCES

- [1] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan, "A theory of learning from different domains," *Machine Learning*, vol. 79, no. 1-2, 2010.
- [2] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *CVPR*, 2012.
- [3] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell, "Adapting visual category models to new domains," in *ECCV*, 2010.
- [4] Brian Kulis, Kate Saenko, and Trevor Darrell, "What you saw is not what you get: Domain adaptation using asymmetric kernel transforms," in *CVPR*, 2011.
- [5] Judy Hoffman, Sergio Guadarrama, Eric Tzeng, Ronghang Hu, Jeff Donahue, Ross B. Girshick, Trevor Darrell, and Kate Saenko, "LSDA: large scale detection through adaptation," in *NIPS*, 2014.



**Fig. 2.** Parameter sensitivity analysis and convergence study of the proposed DE approach.

- [6] Lixin Duan, Dong Xu, and Shih-Fu Chang, “Exploiting web images for event recognition in consumer videos: A multiple source domain adaptation approach,” in *CVPR*, 2012.
- [7] De-An Huang and Yu-Chiang Frank Wang, “Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition,” in *ICCV*, 2013.
- [8] Sinno Jialin Pan, James T. Kwok, and Qiang Yang, “Transfer learning via dimensionality reduction,” in *AAAI*, 2008.
- [9] Mahsa Baktashmotlagh, Mehrtash Tafazzoli Harandi, Brian C. Lovell, and Mathieu Salzmann, “Unsupervised domain adaptation by domain invariant projection,” in *ICCV*, 2013.
- [10] Hal Daumé III, “Frustratingly easy domain adaptation,” in *ACL*, 2007.
- [11] Wen Li, Lixin Duan, Dong Xu, and Ivor W. Tsang, “Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 6, pp. 1134–1148, 2014.
- [12] Jie Ni, Qiang Qiu, and Rama Chellappa, “Subspace interpolation via dictionary learning for unsupervised domain adaptation,” in *CVPR*, 2013.
- [13] Lixin Duan, Ivor W. Tsang, and Dong Xu, “Domain transfer multiple kernel learning,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 3, pp. 465–479, 2012.
- [14] Yuan Shi and Fei Sha, “Information-theoretical learning of discriminative clusters for unsupervised domain adaptation,” in *ICML*, 2012.
- [15] Raghuraman Gopalan, Ruonan Li, and Rama Chellappa, “Domain adaptation for object recognition: An unsupervised approach,” in *ICCV*, 2011.
- [16] Le Shu, Tianyang Ma, and Longin Jan Latecki, “Locality preserving projection for domain adaptation with multi-objective learning,” in *AAAI*, 2014.
- [17] Mahsa Baktashmotlagh, Mehrtash Tafazzoli Harandi, Brian C. Lovell, and Mathieu Salzmann, “Domain adaptation on the statistical manifold,” in *CVPR*, 2014.
- [18] Mingsheng Long, Guiguang Ding, Jianmin Wang, Jiguang Sun, Yuchen Guo, and Philip S. Yu, “Transfer sparse coding for robust image representation,” in *CVPR*, 2013.
- [19] Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiguang Sun, and Philip S. Yu, “Transfer joint matching for unsupervised domain adaptation,” in *CVPR*, 2014.
- [20] Meng Yang, Lei Zhang, Xiangchu Feng, and David Zhang, “Fisher discrimination dictionary learning for sparse representation,” in *ICCV*, 2011.
- [21] Jonathan J. Hull, “A database for handwritten text recognition research,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 5, pp. 550–554, 1994.
- [22] Yann Lecun, Lon Bottou, Yoshua Bengio, and Patrick Haffner, “Gradient-based learning applied to document recognition,” in *Proceedings of the IEEE*, 1998, pp. 2278–2324.
- [23] Ralph Gross, Iain Matthews, Jeffrey F. Cohn, Takeo Kanade, and Simon Baker, “Multi-pie,” *Image Vision Comput.*, vol. 28, no. 5, pp. 807–813, 2010.
- [24] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars, “Unsupervised visual domain adaptation using subspace alignment,” in *ICCV*, 2013.
- [25] Lihi Zelnik-Manor and Pietro Perona, “Self-tuning spectral clustering,” in *NIPS*, 2004.