Distribution Regularized Nonnegative Matrix Factorization for Transfer Visual Feature Learning

Yuchen Guo
Tsinghua Univeristy
Beijing, China
yuchen.w.guo@gmail.com

Guiguang Ding
Tsinghua Univeristy
Beijing, China
dinggg@tsinghua.edu.cn

Qiang Liu
Tsinghua Univeristy
Beijing, China
liuqiang@tsinghua.edu.cn

ABSTRACT

Transfer visual feature learning (TVFL), which learns compact representations for images such that we can build accurate classifier for target domain by leveraging rich labeled data in the source domain, has attracted increasingly attention recently. Previous methods mainly focus on reducing the distribution difference between domains but ignore the intrinsic hidden semantics in data. In this paper, we put forward a novel method for TVFL, called **D**istribution **R**egularized **N**onnegative **M**atrix **F**actorization (DRNMF). Specifically, we employ Nonnegative Matrix Factorization (NMF) to uncover the intrinsic information in visual data, and regularize it with geometrical distribution, marginal probability distribution and conditional probability distribution. Thus, DRNMF can discover the intrinsic information, preserve the manifold structure and reducing both marginal and conditional probability distribution difference simultaneously, which all perspectives above are important for TVFL. We also propose an effective and efficient algorithm for the optimization of DRNMF and theoretically prove the convergence. Extensive experiments on three types of cross-domain image classification tasks in comparison with several state-of-the-art methods demonstrate the superiority of our DRNMF, which validates its effectiveness.

General Terms

Algorithms, Performance, Experimentation

Keywords

Nonnegative Matrix Factorization, Transfer Learning, Geometrical Distribution, Probability Distribution

1. INTRODUCTION

Given a new target domain for classification, it's sometimes burdensome and difficult, even if not impossible, to obtain sufficient labeled data to train accurate classifiers [23, 26]. Fortunately, there are always related source domains where labeled data is abundant and we can use the knowledge in source domains to help build accurate classifiers for target domain, which is termed in literatures as *transfer learning* [24] who has shown promising results in image classification [14, 29], object recognition [1, 11], feature learning [13, 15] and retrieval [7]. Because of the probability distribution difference between domains, previous works on transfer learning mainly focus on learning *transfer features* such that the inter-domain probability difference can be reduced by preserving statistical properties [13, 25], geometric structure [8, 27] shared by different domains, or explicitly minimizing the pre-defined distance measures [21, 22, 23, 28, 30]. Though the learned transfer features can alleviate the probability distribution difference between domains such that we can train standard classifiers like Logistic Regression and SVM on them, they can't capture the intrinsic hidden semantics of image data which is also a very important perspective for building effective classifiers.

Actually, one may always hope to construct compact low-dimensional representations which can also uncover the intrinsic hidden semantics of image data. Several feature learning techniques have been proposed for image data to achieve this goal. Among them, Nonnegative Matrix Factorization (NMF) [16, 17] has recently attracted increasingly attention because the nonnegative constraints can lead to *parts-based* representation for images because only additive, not subtractive, combinations of basis vectors are allowed. And there is psychological and physiological evidence for parts-based representations in human brain [19]. NMF has been proved to be an effective tool to extract intrinsic hidden semantics and applied to several tasks like face recognition [18] and image clustering [3, 5]. But when the labeled and unlabeled images are drawn from different probability distributions, NMF will be greatly challenged because it may generate quite different representations for images drawn from different probability distributions even if they share the same semantic label, e.g., images of one object but collected from different illumination conditions. Fortunately, compared to other feature learning techniques like Sparse Coding which always leads to high-dimensional representation with over-complete basis, it's more natural to incorporate transfer learning to NMF since NMF leads to features with low dimensionality which is more favored than the high one in transfer learning who always looks for low-dimensional shared subspace for domains [21, 23, 28].

Motivated by recent works in transfer learning and NMF, in this paper we propose a novel method for transfer visual feature learning, called Distribution Regularized Non-

ICMR '15 June 23 - 26, 2015, Shanghai, China

http://dx.doi.org/10.1145/2671188.2749316.

negative Matrix Factorization (DRNMF). Specifically, it's build on NMF, and regularized by geometrical distribution, marginal probability distribution and conditional probability distribution. Thus it can simultaneously uncover the intrinsic hidden semantics, preserve local geometric structure, reduce difference of both marginal and conditional probability distributions. By utilizing the power of NMF, DRNMF can learn more effective visual features to build accurate cross-domain classifiers. Meanwhile, DRNMF can also be regarded as an enhanced and extended version of NMF such that NMF can handle the probability distribution difference between datasets, which is also very common in real world. In summary, this paper makes some contributions as below.

- We propose a novel DRNMF for transfer visual feature learning. It can learn more effective compact features for cross-domain image classification task by simultaneously extracting intrinsic hidden semantics, preserving geometric structure, and reducing marginal as well as conditional distribution difference among domains.
- Conventional NMF is extended by DRNMF such that it can handle the probability distribution difference in data while previous works on NMF almost ignore this.
- We put forward an effective and efficient iterative algorithm with multiplication updating rules for the optimization problem of DRNMF, and give the theoretical analysis on the convergency property of this algorithm.
- We conduct extensive experiments on three types of cross-domain image classification problems. The experiment results demonstrate that DRNMF can significantly outperform several state-of-the-art related methods, validating the effectiveness of our DRNMF.

The rest of this paper is organized as follows. In Section 2, some important related works are reviewed. The the proposed DRNMF is introduced in detail in Section 3. The theoretical analysis is shown in Section 4. Then we present experiment in Section 5 and draw conclusion in Section 6.

2. RELATED WORK

Maximum Mean Discrepancy (MMD)[9] is widely used as an nonparametric distance measure between probability distributions. It will asymptotically approach to zero if and only if the two distributions are identical. By minimizing MMD of features, the probability distribution difference of original features in different domains can be alleviated. Specifically, Transfer Component Analysis (TCA) [23] aims to minimize the MMD of marginal probability distribution between domains and the reconstruction error of data simultaneously based on PCA. And Joint Distribution Adaptation (JDA) [21] extends TCA by further minimizing the MMD of conditional probability distribution between domains. Though the probability distribution difference between domains can be reduced, these methods can't discover the intrinsic hidden semantics, as mentioned above, which is also a key difference between their methods and DRNMF who focuses on learning effective image representations by building an adaptive model based on NMF.

We also noticed that some efforts have been made to incorporate effective feature learning methods into transfer learning such that the learned transfer visual features can reduce the probability distribution difference while discover intrinsic hidden semantics [20, 26]. These methods adopt Sparse Coding (SC) for feature learning. But SC always requires over-complete basis such that it's difficult to directly apply SC to high-dimensional image data and the learned features are not compact and low-dimensional. And they need to solve an ℓ_1-regularized least square problem which is highly inefficient compared to NMF. Furthermore, kernel density estimation technique is adopted in [26] which is more restricted than DRNMF and is prone to be over-fitting, while conditional probability distribution is ignored in [20].

Table 1: Notations and descriptions in this paper

Notation	Description	Notation	Description
D_s	source domain	D_t	target domain
n_s	#source images	n_t	#target iamges
m	#original features	C	#classes
k	#basis vectors	p	#NN
α	graph para.	β	MMD para.
$\mathbf{X}$	input matrix	$\mathbf{U}$	basis matrix
$\mathbf{V}$	transfer features	$\mathbf{W}$	NN matrix
$\mathbf{L}$	graph Lap. matrix	$\mathbf{M}_c$	MMD matrix

3. THE PROPOSED METHOD

3.1 Problem Definition

Following the conventional definition for transfer feature learning, such as in [21], our problem is defined as follows,

PROBLEM 1. *Given a labeled dataset in source domain $\mathcal{D}_s = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_{n_s}, y_{n_s})\}$, and an unlabeled dataset in target domain $\mathcal{D}_t = \{\mathbf{x}_{n_s+1}, ..., \mathbf{x}_{n_s+n_t}\}$, where the marginal and conditional probability distributions are both different, i.e., $P_s(\mathbf{x}_s) \neq P_t(\mathbf{x}_t)$ and $P_s(y_s|\mathbf{x}_s) \neq P_t(y_t|\mathbf{x}_t)$, our goal is to learn a new representation such that classifiers trained on labeled dataset can work robustly on unlabeled dataset.*

where n_s and n_t are the number of images in source and target domain respectively. Totally, we can first define $\mathbf{X} = [x_1, ..., x_{n_s+n_t}] \in \mathbb{R}^{m\times(n_s+n_t)}$ as the nonnegative data matrix where each data is represented by an m-dimensional vector. Our goal is to learn a basis matrix $\mathbf{U} \in \mathbb{R}^{m\times k}$ and corresponding k-dimensional nonnegative transfer features $\mathbf{V} = [\mathbf{v}_1, ..., \mathbf{v}_{n_s+n_t}] \in \mathbb{R}^{k\times(n_s+n_t)}$. Now we can train any supervised classifiers on $\{(\mathbf{v}_1, y_1), ..., (\mathbf{v}_{n_s}, y_{n_s})\}$, and then classify $\{\mathbf{v}_{n_s+1}, ..., \mathbf{v}_{n_s+n_t}\}$ with obtained classifiers.

3.2 Objective Function

3.2.1 Nonnegative Matrix Factorization

Given a nonnegative data matrix $\mathbf{X}$, NMF aims to find two nonnegative matrices $\mathbf{U}$ as basis of latent space and the corresponding coordinates $\mathbf{V}$ which can well approximate the original data matrix $\mathbf{X}$. And the quality of approximation is always measured by the squared loss function. Therefore the objective function of NMF can be written as below [16],

$$\mathcal{O}_{\text{NMF}} = \|\mathbf{X} - \mathbf{U}\mathbf{V}\|_F^2 \quad \text{s.t.} \quad \mathbf{U}, \mathbf{V} \geq 0 \tag{1}$$

where $\|\cdot\|_F$ is the Frobenius norm of matrix. NMF has been proved to be a powerful tool to learn compact low-dimensional representations and uncover the intrinsic hidden semantics of image data because the nonnegative constraints may lead to parts-based representations for image data which has psychological and physiological evidence in human brain [19]. Generally, minimizing Eq. (1) can be

effectively and efficiently achieved by utilizing an iterative strategy with the following multiplicative updating rules [17]

$$u_{il} \leftarrow u_{il}\frac{(\mathbf{X}\mathbf{V}^T)_{il}}{(\mathbf{U}\mathbf{V}\mathbf{V}^T)_{il}}, \quad v_{lj} \leftarrow v_{lj}\frac{(\mathbf{U}^T\mathbf{X})_{lj}}{(\mathbf{U}^T\mathbf{U}\mathbf{V})_{lj}} \tag{2}$$

3.2.2 Geometrical Distribution Regularization

In real world, natural data always lies on a low-dimensional manifold embedded in high-dimensional ambient space. However, conventional NMF ignores this *geometrical distribution* underlying the image data. In [3], an extension of NMF called Graph Regularized NMF (GNMF) is proposed to further explore the geometrical distribution of data. Specifically, the geometrical distribution is exploited by preserving the locality of data based on the locality invariant idea [12], i.e., the nearby points in original space should be close to each other in low-dimensional latent space. Firstly we can construct a p (such as 5) nearest neighbor matrix as

$$W_{ij} = \begin{cases} 1, & \text{if } \mathbf{x}_i \in \mathcal{N}(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in \mathcal{N}(\mathbf{x}_i) \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where $\mathcal{N}(\mathbf{x}_i)$ denotes the p nearest neighbor of $\mathbf{x}_i$. Then we can define a diagonal degree matrix with diagonal element $D_{ii} = \sum_{j=1}^{n} W_{ij}$ and the graph Laplacian matrix $\mathbf{L} = \mathbf{D} - \mathbf{W}$. So the locality invariant idea can be formulate as minimizing $\frac{1}{2}\sum_{i,j=1}^{n} W_{ij}\|\mathbf{v}_i - \mathbf{v}_j\|_F^2 = tr(\mathbf{V}\mathbf{L}\mathbf{V}^T)$. By incorporating this geometrical distribution regularization (or termed as graph regularization in [3]) into NMF, we have the objective function of GNMF written as below,

$$\mathcal{O}_{\text{GNMF}} = \|\mathbf{X} - \mathbf{U}\mathbf{V}\|_F^2 + \alpha tr(\mathbf{V}\mathbf{L}\mathbf{V}^T) \text{ s.t. } \mathbf{U}, \mathbf{V} \geq 0 \tag{4}$$

where α is the regularization parameter to control the weight of the regularization. Actually, GNMF has shown promising result on single-domain problems because it's able to uncover the intrinsic hidden semantics and preserve the geometrical distribution of data. However, when applying it directly to cross-domain problems, the performance of GNMF may degrade severely because it ignores the probability distribution difference between domains. But it lays the foundation of our DRNMF as a feature learning method.

3.2.3 Probability Distribution Regularization

In the k-dimensional representation learned by GNMF, the distribution difference between domains is still significantly large. Thus we need to reduce the distribution difference by explicitly minimize some predefined distance measure during the feature learning procedure. In this paper we follow [9, 22, 23] and utilize a nonparametric distance measure, Maximum Mean Discrepancy (MMD), to compare different distributions, which is defined as the distance between the sample means of source and target domains as

$$\|\frac{1}{n_s}\sum_{i=1}^{n_s}\mathbf{v}_i - \frac{1}{n_t}\sum_{j=n_s+1}^{n_s+n_t}\mathbf{v}_j\|_F^2 = tr(\mathbf{V}\mathbf{M}_0\mathbf{V}^T) \tag{5}$$

where $\mathbf{M}_0$ is the MMD matrix which is defined as follows,

$$(M_0)_{ij} = \begin{cases} \frac{1}{n_s n_s}, & \text{if } \mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}_s \\ \frac{1}{n_t n_t}, & \text{if } \mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}_t \\ \frac{-1}{n_s n_t}, & \text{otherwise} \end{cases} \tag{6}$$

By minimizing Eq. (5), we can reduce the **marginal** probability distribution difference between domains in learned representations like in TCA, i.e., we have $P_s(\mathbf{v}_s) \approx P_t(\mathbf{v}_t)$.

Furthermore, the conditional probability distribution is also important for transfer feature learning, especially when the learned features are utilized in classification task. Some works have been proposed to match the conditional probability distributions [2, 4, 26, 30]. But they all require some labeled data in target domain. Thus they can't be applied to our problem. Instead, we follow the idea in [21] where *pseudo* labels of target domain data predicted by base classifier which is trained on the labeled source domain data are exploited to match the conditional probability distributions.

Actually, directly matching the sample-conditional probability distribution $P(y|\mathbf{v})$ is infeasible because this posterior probability is very difficult to estimate. Instead, we can match the class-conditional probability distribution $P(\mathbf{v}|y)$, i.e., we hope the learned representations can satisfy $P_s(\mathbf{v}_s|y = c) \approx P_t(\mathbf{v}_t|y = c)$ for $\forall c \in \{1, ..., C\}$, where C is the number of class. Besides the difficulty to estimate sample-conditional probability distribution, here is another reason why we can utilize class-conditional probability distribution. If Eq. (5) is effectively minimized, we can obtain $P_s(\mathbf{v}_s) \approx P_t(\mathbf{v}_t)$. Under the ultimate situation, we have $P_s(\mathbf{v}_s) = P_t(\mathbf{v}_t)$. Then by assuming $P(y)$ is consistent in both domains, i.e., $P_s(y) = P_t(y)$, we can get $P_s(y|\mathbf{v}_s) - P_t(y|\mathbf{v}_t) \propto P_s(\mathbf{v}_s|y) - P_t(\mathbf{v}_t|y)$ based on Bayesian formula $P(y|\mathbf{v}) = P(\mathbf{v}|y)P(y)/P(\mathbf{v})$. Therefore matching the class-conditional probability distribution $P(\mathbf{v}|y)$ is equivalent to matching the sample-conditional probability distribution $P(y|\mathbf{v})$. Thus, given the true labels of labeled source domain data and pseudo labels of unlabeled target domain data, we can match the class-conditional probability distributions of different domains by minimizing the MMD for images in each class c as follows,

$$\|\frac{1}{n_s^c}\sum_{\mathbf{x}_i \in \mathcal{D}_s^c}\mathbf{v}_i - \frac{1}{n_t^c}\sum_{\mathbf{x}_j \in \mathcal{D}_t^c}\mathbf{v}_j\|_F^2 = tr(\mathbf{V}\mathbf{M}_c\mathbf{V}^T) \tag{7}$$

where $\mathcal{D}_s^c$ is the subset of $\mathcal{D}_s$ in which the **true** labels of images are c and $n_s^c = |\mathcal{D}_s^c|$. And $\mathcal{D}_t^c$ is the subset of $\mathcal{D}_t$ in which the **pseudo** labels of images are c and $n_t^c = |\mathcal{D}_t^c|$. And the MMD matrix $\mathbf{M}_c$ for class c is defined as follows,

$$(M_c)_{ij} = \begin{cases} \frac{1}{n_s^c n_s^c}, & \text{if } \mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}_s^c \\ \frac{1}{n_t^c n_t^c}, & \text{if } \mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}_t^c \\ \frac{-1}{n_s n_t}, & \text{if } \mathbf{x}_i(\mathbf{x}_j) \in \mathcal{D}_s^c, \mathbf{x}_j(\mathbf{x}_i) \in \mathcal{D}_t^c \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

Then by minimizing Eq. (7), the conditional probability distribution difference between domains for each class can be effectively reduced in the learned feature representations.

Here, the pseudo labels for target domain can be generated by applying a classifier trained on the labeled source domain to the target domain. Generally, 1 nearest neighbor (1NN) classifier can lead to satisfactory result. However, many of initial pseudo labels are incorrect because the conditional probability distributions aren't correctly matched at first. To address this issue, we can conduct the feature learning step and pseudo label generating step iteratively thus we can expect to obtain more correct pseudo labels with more iterations, leading to better matched conditional probability distributions and vice versa. In our experiment, we find out that satisfactory result can be achieved in 5 to 10 iterations.

3.2.4 Overall Objective Function

We first define the overall MMD matrix as $\mathbf{M} = \sum_{c=0}^{C}\mathbf{M}_c$.

Algorithm 1 Learning Transfer Features by DRNMF

Input:
Image data $\mathbf{X} = [\mathbf{X}_s, \mathbf{X}_t]$, true labels $\mathbf{Y}_s$, #basis vectors k, graph Lap. $\mathbf{L}$, regul. para. α, β, #iteration T

Output:
The learned basis $\mathbf{U}$ and transfer features $\mathbf{V}$.

1: Construct $\mathbf{M}_0$ by Eq. (6), and $\mathbf{M}_c = \mathbf{0}$, $c = 1, ..., C$
2: **for** $t = 1 : T$ **do**
3: Construct $\mathbf{R} = \alpha\mathbf{L} + \beta \sum_{c=0}^{C} \mathbf{M}_c$.
4: Initialize $\mathbf{U}$ and $\mathbf{V}$ by randomization.
5: **repeat**
6: Update $\mathbf{U}$ by Eq. (16).
7: Update $\mathbf{V}$ by Eq. (17).
8: **until** Convergence
9: Update pseudo labels $\hat{\mathbf{Y}}_t$ by 1NN classifier.
10: Construct $\mathbf{M}_c$ by Eq. (8), $c = 1, ..., C$.
11: **end for**
12: Return basis $\mathbf{U}$ and transfer features $\mathbf{V}$.

Then by combining Eq. (4), Eq. (5) and Eq. (7), we can obtain the overall objective function of DRNMF as follows,

$$\mathcal{O} = \|\mathbf{X} - \mathbf{UV}\|_F^2 + tr(\mathbf{VRV}^T) \text{ s.t. } \mathbf{U}, \mathbf{V} \geq 0 \quad (9)$$

where $\mathbf{R} = \alpha\mathbf{L} + \beta\mathbf{M}$ is the overall distribution regularization matrix which combines the geometric distribution regularization, marginal probability distribution regularization and conditional probability distribution regularization. $\alpha > 0$ and $\beta > 0$ are the regularization parameters which trade off the weight of geometrical distribution regularization and probability distribution regularization respectively.

Actually, every term in Eq. (9) is indispensable for transfer feature learning. NMF uncovers the intrinsic hidden semantics of data leading to more meaningful representation, which is an important difference and improvement compared to several transfer **feature** learning methods like TCA and JDA. The geometrical distribution regularization explores the low-dimensional geometric structure underlying the data, which can promote the effectiveness of the learned representations. The probability regularization can reduce both the marginal and conditional probability distribution difference between domains such that we can build accurate classifiers on the learned representations, which is also a basic requirement for a **transfer** feature learning method.

3.3 Optimization Algorithm

The optimization problem of minimizing Eq. (9) is not convex with $\mathbf{U}$ and $\mathbf{V}$ together. Fortunately, it's convex with respect to any one while fixing the other. Therefore we can utilize the following iterative strategy by updating one while fixing the other which will achieve the local minima.

Based on the following matrix properties, $\|\mathbf{A}\|_F^2 = tr(\mathbf{AA}^T)$, $tr(\mathbf{AB}) = tr(\mathbf{BA})$ and $tr(\mathbf{A}) = tr(\mathbf{A}^T)$, the objective function of DRNMF in Eq. (9) can be rewritten as follows,

$$\begin{aligned}\mathcal{O}_1 =& tr(\mathbf{XX}^T) - 2tr(\mathbf{XV}^T\mathbf{U}^T) + tr(\mathbf{VRV}^T) \\ &+ tr(\mathbf{UVV}^T\mathbf{U}^T), \text{ s.t. } \mathbf{U}, \mathbf{V} \geq 0\end{aligned} \quad (10)$$

Now denote ψ_{il} and ϕ_{lj} as the Lagrange multiplier for constraint $u_{il} \geq 0$ and $v_{lj} \geq 0$ respectively, and $\mathbf{\Psi} = [\psi_{il}]$ and $\mathbf{\Phi} = [\phi_{lj}]$. Then we can rewrite the Lagrange $\mathcal{L}$ as follows,

$$\mathcal{L} = \mathcal{O}_1 + tr(\mathbf{\Psi}\mathbf{U}^T) + tr(\mathbf{\Phi}\mathbf{V}^T) \quad (11)$$

The partial derivatives of $\mathcal{L}$ with respect to $\mathbf{U}$ and $\mathbf{V}$ are

$$\frac{\partial \mathcal{L}}{\partial \mathbf{U}} = -2\mathbf{XV}^T + 2\mathbf{UVV}^T + \mathbf{\Psi} \quad (12)$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{V}} = -2\mathbf{U}^T\mathbf{X} + 2\mathbf{U}^T\mathbf{UV} + 2\mathbf{VR} + \mathbf{\Phi} \quad (13)$$

Then by using the Karush-Kuhn-Tucker conditions, that is, $\psi_{il}u_{il} = 0$ and $\phi_{lj}v_{lj} = 0$, we get the following equations,

$$-(\mathbf{XV}^T)_{il}u_{il} + (\mathbf{UVV}^T)_{il}u_{il} = 0 \quad (14)$$

$$-(\mathbf{U}^T\mathbf{X})_{lj}v_{lj} + (\mathbf{U}^T\mathbf{UV})_{jl}v_{jl} + (\mathbf{VR})_{jl}v_{jl} = 0 \quad (15)$$

Then we obtain the following multiplicative updating rules:

$$u_{il} \leftarrow u_{il}\frac{(\mathbf{XV}^T)il}{(\mathbf{UVV}^T)_{il}} \quad (16)$$

$$v_{lj} \leftarrow v_{lj}\frac{(\mathbf{U}^T\mathbf{X} + \mathbf{VR}^-)_{lj}}{(\mathbf{U}^T\mathbf{UV} + \mathbf{VR}^+)_{lj}} \quad (17)$$

where $\mathbf{R}^+ = \frac{1}{2}(|\mathbf{R}| + \mathbf{R})$ is the positive part of $\mathbf{R}$ and $\mathbf{R}^- = \frac{1}{2}(|\mathbf{R}| - \mathbf{R})$ is the negative part of $\mathbf{R}$. Applying Eq. (16) and Eq. (17) iteratively, $\mathcal{O}$ in Eq. (9) can reach local minima. Then based on the learned transfer representation $\mathbf{V}$, we can update the pseudo labels of target domain data by a new 1NN classifier trained on source domain data. We can repeat these two steps for T times for the final representation. Generally, setting $T = 10$ can result in satisfactory performance and guarantee efficiency. We summarize the overall optimization algorithm for DRNMF in Algorithm 1.

4. THEORETICAL ANALYSIS

4.1 Proof of Convergence

Actually, it's important to note that there is no theoretical evidence that the outer iteration can converge because we have no labeled data and prior knowledge in target domain. And because of the random initialization and the local minima, we may obtain different pseudo labels even with the same $\mathbf{L}$ and $\mathbf{M}$. But fortunately, we find out in our experiment that we can usually obtain more correct pseudo labels in each iteration because the conditional probability distribution is indeed better matched. And it will finally achieve a relatively stable performance, generally within about 10 iterations. Thus this EM-like iterative strategy is always effective in reality, which is also validated in our experiment.

But we need to guarantee the convergence of the inner iteration, i.e., the feature learning step (line 5 to 8). Otherwise we can't obtain effective and meaningful representations. The convergence is guaranteed by Theorem 1 below.

THEOREM 1. *The objective function $\mathcal{O}$ in Eq. (9) is nonincreasing under updating rules in Eq. (16) and Eq. (17).*

The objective function is obviously lower-bounded by zero. Thus $\mathcal{O}$ will certainly converge if it's nonincreasing. To prove Theorem 1, we need to show $\mathcal{O}$ is nonincreasing under Eq. (16) and Eq. (17). Actually, The part of $\mathcal{O}$ which is only relevant to $\mathbf{U}$ is the same as NMF and the updating rule in Eq. (16) is also exactly the same as in original NMF. Therefore $\mathcal{O}$ is nonincreasing under Eq. (16), whose detailed proof can be found in [17]. Now we need to show $\mathcal{O}$ is nonincreasing under Eq. (17). Following the proof in [6], firstly we introduce the definition of *auxiliary function*

DEFINITION 1. $A(v, v')$ is an *auxiliary function* for $B(v)$ if

$$A(v, v') \geq B(v), \quad A(v, v) = B(v) \tag{18}$$

are satisfied.

Then we need to introduce the following important lemma,

LEMMA 1. If $A(v, v')$ is an auxiliary function of $B(v)$, then $B(v)$ will be nonincreasing under the following update

$$v^{(t+1)} = \arg\min_{v} A(v, v^{(t)}) \tag{19}$$

PROOF PROOF OF LEMMA 1.

$$B(v^{(t+1)}) \leq A(v^{(t+1)}, v^{(t)}) \leq A(v^{(t)}, v^{(t)}) = B(v^{(t)})$$

□

Now we need to show the updating rule for $\mathbf{V}$ in Eq. (17) is exactly the update in Eq. (19) with a proper auxiliary function. Let B_{ab} denote the the part of $\mathcal{O}$ that is only relevant to v_{ab}. Then the second-order partial derivative of B_{ab} is as

$$B_{ab}^{''} = 2(\mathbf{U}^T\mathbf{U})_{aa} + 2\mathbf{R}_{bb} \tag{20}$$

And it is quite easy to verify the following two inequalities,

$$(\mathbf{U}^T\mathbf{U}\mathbf{V})_{ab} = \sum_{i=1}^{k} (\mathbf{U}^T\mathbf{U})_{ak} v_{kb}^{(t)} \geq (\mathbf{U}^T\mathbf{U})_{aa} v_{ab}^{(t)} \tag{21}$$

$$(\mathbf{V}\mathbf{R}^+)_{ab} = \sum_{j=1}^{n} v_{aj}^{(t)} \mathbf{R}_{jb}^+ \geq v_{ab}^{(t)} \mathbf{R}_{bb}^+ \geq v_{ab}^{(t)} \mathbf{R}_{bb} \tag{22}$$

LEMMA 2. The function

$$\begin{aligned} A(v, v_{ab}^{(t)}) =& B_{ab}(v_{ab}^{(t)}) + B_{ab}^{'}(v_{ab}^{(t)})(v - v_{ab}^{(t)}) \\ &+ \frac{(\mathbf{U}^T\mathbf{U}\mathbf{V} + \mathbf{V}\mathbf{R}^+)_{ab}}{v_{ab}^{(t)}}(v - v_{ab}^{(t)})^2 \end{aligned} \tag{23}$$

is an auxiliary function for $B_{ab}(v)$.

PROOF PROOF TO LEMMA 2. It's obvious that $A(v, v) = B_{ab}(v)$. Now we need to show $A(v, v_{ab}^{(t)}) \geq B_{ab}(v)$. Here we compare the Taylor series expansion of $B_{ab}(v)$ at $v_{ab}^{(t)}$ as

$$\begin{aligned} B_{ab}(v) =& B_{ab}(v_{ab}^{(t)}) + B_{ab}^{'}(v_{ab}^{(t)})(v - v_{ab}^{(t)}) \\ &+ \frac{1}{2} B_{ab}^{''}(v_{ab}^{(t)})(v - v_{ab}^{(t)})^2 \end{aligned} \tag{24}$$

Based on the definition in Eq. (20) and Eq. (23), and two important inequalities presented in Eq. (21) and Eq. (22), it's quite straightforward to verify that $A(v, v_{ab}^{(t)}) \geq B_{ab}(v)$. □

PROOF PROOF OF THEOREM 1. We can replace $A(v, v_{ab}^{(t)})$ in Eq. (19) by Eq. (23), which result in the following rule,

$$\begin{aligned} v_{ab}^{(t+1)} &= v_{ab}^{(t)} - v_{ab}^{(t)} \frac{B_{ab}^{'}(v_{ab}^{(t)})}{2(\mathbf{U}^T\mathbf{U}\mathbf{V} + \mathbf{V}\mathbf{R}^+)_{ab}} \\ &= v_{ab}^{(t)} \frac{(\mathbf{U}^T\mathbf{X} + \mathbf{V}\mathbf{R}^-)_{ab}}{(\mathbf{U}^T\mathbf{U}\mathbf{V} + \mathbf{V}\mathbf{R}^+)_{ab}} \end{aligned} \tag{25}$$

which is identical to Eq. (17). Because $A(v, v_{ab}^{(t)})$ is an auxiliary function of B_{ab}, B_{ab} is nonincreasing under this updating rule. Therefore $\mathcal{O}$ is nonincreasing under Eq. (17). □

Table 2: Statistics of benchmark datasets

Dataset	Type	#Img	#Fea	#Class
MNIST	Digit	1,000	256	10
USPS	Digit	1,000	256	10
Semeion	Digit	1,000	256	10
COIL1,2	Object	720	1024	20
PIE1,...,5	Face	1,360	1024	68

Figure 1: USPS, MNIST, Semeion, COIL, and PIE

4.2 Complexity Analysis

Suppose we have T outer iterations and t inner iterations, then the overall time complexity for Algorithm 1 is $\mathcal{O}(Tt(mnk+(m+n)(k^2+k)+n^2k))$. We can observe that the complexity is linear to the feature dimension m. Therefore the optimization algorithm is efficient for high-dimensional feature which is very common for images data.

5. EXPERIMENT AND DISCUSSION

5.1 Experiment Setting

5.1.1 Benchmark Datasets

We conduct extensive experiment on three types of benchmark datasets for cross-domain image classification task, MNIST[1]+USPS[2] +Semeion[3], COIL[4] and PIE[5]. We construct several subsets from them for implementation efficiency. The statistics of these datasets are summarized in Table 2. We also present several sample images in Figure 1.

USPS, MNIST and Semeion are three widely used handwritten digit datasets. To speed up experiments, we select $1,000$ images from each dataset, which consists of 10 classes and 100 images per class. We rescale all images to size 16×16 and represent each image by a vector encoding the gray-scale pixel values. From Figure 1, we can observe that these three datasets follow quite different distributions. Thus we can construct 6 cross-domain classification tasks, e.g., *USPS vs MNIST* in which we use USPS as the labeled source domain and MNIST as the unlabeled target domain.

COIL contains 1,440 images of 20 objects taken from different degrees. We split it into two subsets, COIL1 containing images taken in directions of $[0°, 85°] \cup [180°, 265°]$ and COIL2 in directions of $[90°, 175°] \cup [270°, 355°]$. Both contain 20 classes and 36 images per class. Thus COIL1 and COIL2 follow different distributions. Each image is represented by a 1,024-dimension vector encoding the gray-scale

[1]http://yann.lecun.com/exdb/mnist/
[2]1http://www-i6.informatik.rwth-aachen.de
[3]https://archive.ics.uci.edu/ml/datasets
[4]http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php
[5]https://www.ri.cmu.edu/research-project-detail.html

Table 3: Cross-domain Image Classification Accuracy (%)

Dataset	LR	PCA	NMF	GNMF	TCA	TSL	GFK	TSC	JDA	Ours
MNIST vs USPS	41.30	41.40	47.08	50.36	50.20	66.40	64.28	67.84	71.30	**88.83**
USPS vs MNIST	34.30	34.30	36.64	39.71	40.00	46.10	38.91	46.85	53.30	**56.07**
MNIST vs Semeion	7.50	6.30	10.19	10.41	7.10	16.20	11.37	17.85	19.10	**30.63**
Semeion vs MNIST	10.40	10.50	8.14	11.74	7.90	17.80	13.80	20.57	28.20	**29.49**
USPS vs Semeion	38.90	38.40	34.31	36.69	41.60	48.00	39.71	41.32	55.30	**58.11**
Semeion vs USPS	42.30	42.00	36.97	41.72	55.30	58.90	52.17	65.41	69.10	**76.13**
COIL1 vs COIL2	76.11	75.56	74.33	77.31	78.47	82.84	76.31	85.12	88.75	**94.31**
COIL2 vs COIL1	74.03	73.75	73.56	75.06	78.61	80.61	77.84	84.33	86.67	**95.00**
PIE1 vs PIE5	21.03	20.15	30.15	30.64	36.25	38.75	32.26	40.31	42.87	**52.13**
PIE2 vs PIE4	64.56	62.35	55.88	59.37	68.75	73.28	63.41	71.85	76.18	**80.02**
PIE3 vs PIE1	29.49	29.19	35.66	38.18	46.83	44.84	40.08	46.84	49.56	**61.10**
PIE4 vs PIE3	67.21	64.48	68.68	70.59	72.21	75.85	70.26	76.21	80.22	**86.69**
PIE5 vs PIE2	31.39	29.63	30.81	32.62	37.57	38.17	39.78	36.04	42.79	**44.12**
Average	40.86	39.65	41.77	43.94	49.38	53.53	47.54	54.13	58.17	**64.93**

pixel values. And we can construct two cross-domain classification tasks, *COIL1 vs COIL2* and *COIL2 vs COIL1*.

PIE face dataset contains images of 68 individuals under different poses, illuminations and expressions. For thoroughly comparison, we construct 5 subsets with each corresponding to a different poses, i.e., PIE1 (left), PIE2 (upward), PIE3 (downward), PIE4 (frontal) and PIE5 (right). Each subset contains 68 classes (individuals) with 20 images per class where each image is represented by a 1,024-dimension feature. So we can construct 20 cross-domain classification tasks, i.e., *PIEi vs PIEj* ($i, j = 1, ..., 5, i \neq j$).

5.1.2 Baseline Methods

The proposed DRNMF is compared to several state-of-the-art related feature learning methods for cross-domain image classification task, including both standard feature learning and transfer feature learning methods below.

- Logistic Regression (LR)
- Principle Component Analysis (PCA) + LR
- Nonnegative Matrix Factorization (NMF) [16] + LR
- Graph Regularized NMF (GNMF) [3] + LR
- Transfer Component Analysis (TCA) [23] + LR
- Transfer Subspace Learning (TSL) [28] + LR
- Geodestic Flow Kernel (GFK) [8] + LR
- Transfer Sparse Coding (TSC) [20] + LR
- Joint Distribution Adaptation (JDA) [21] + LR

We choose LR as the base classifier following the setting in [20]. PCA, NMF and GNMF are standard learning methods that NMF can uncover the intrinsic hidden semantics and GNMF has geometrical distribution regularization. TCA, TSL, GFK, TSC and JDA are transfer learning methods. TCA and TSL only consider the marginal probability distribution while JDA considers both marginal and conditional distributions. They can't uncover the hidden semantics of data and ignore local geometric structure. TSC uncovers the hidden semantics by Sparse Coding while it fails to reduce the conditional probability distribution difference.

5.1.3 Implementation Details

For fair comparison, we fix some unimportant parameters for all experiments as follows. The regularization parameter for LR, i.e., c, is consistently set to 1. And for all feature learning methods, we set k, the dimension of the new representation, to 100. And p for constructing p-NN matrix, is set as 5 for GNMF, TSC and DRNMF.

There are some important model parameters for different baseline methods which may have significant effect on the performance of these methods. For meaningful comparison, we empirically search the parameter space for each baseline methods and the *best* result for each baseline method in each experiment is reported. Specifically, the graph regularization parameter in GNMF and TSC, is chosen from $\{0.001, 0.005, 0.01, 0.05, ..., 100\}$. The adaptation regularization parameter for TSL, TCA and JDA is empirically searched in $\{0.01, 0.1, 1, 10, 100\}$. And the MMD regularization and the sparsity regularization parameters in TSC are selected from $\{10^0, ..., 10^6\}$ and $\{0.001, 0.005, ..., 10\}$ respectively.

There are two important parameters for DRNMF, i.e., the geometrical distribution regularization parameter α and the probability distribution regularization parameter β. In the coming section, we conduct empirical analysis on parameter sensitivity which demonstrates that DRNMF can achieve superior and stable performance under a wide range of parameter values for α and β. When comparing with baseline methods, we set 1) $\alpha = 0.1$ and $\beta = 100$ for digit and object datasets, and 2) $\alpha = 0.01$ and $\beta = 10$ for face datasets. And we further set the number of outer iterations, T, to 10.

The classification *Accuracy* on the target domain data is adopted as the evaluation metric, which is widely utilized in related literatures [21, 23, 28]. And it is defined as follows,

$$Accuracy = \frac{|\mathbf{x} : \mathbf{x} \in \mathcal{D}_t \wedge \hat{y}(\mathbf{x}) = y(\mathbf{x})|}{|\mathbf{x} : \mathbf{x} \in \mathcal{D}_t|} \tag{26}$$

where $\mathcal{D}_t$ is the target domain, $\hat{y}(\mathbf{x})$ is the predicted label by LR and $y(\mathbf{x})$ is the true label of $\mathbf{x}$. To remove randomness caused by random initialization, like in NMF and DRNMF, all results reported are the average over 10 runs.

5.2 Experiment Result

5.2.1 Cross-domain Classification Result

The classification results of DRNMF and other 9 baseline methods on all 28 cross-domain datasets are shown in Table 3 and Figure 2. Because of the limit of space, we just list 5 datasets of PIE in Table 3, and all 20 results of PIE are shown in Figure 2. We can observe that DRNMF can consistently and significantly outperform all baseline methods on all three types of tasks, i.e., digit, object and face. The average accuracy of DRNMF on all 28 datasets is **64.93**%, and the improvement compared to best baseline methods, i.e.,

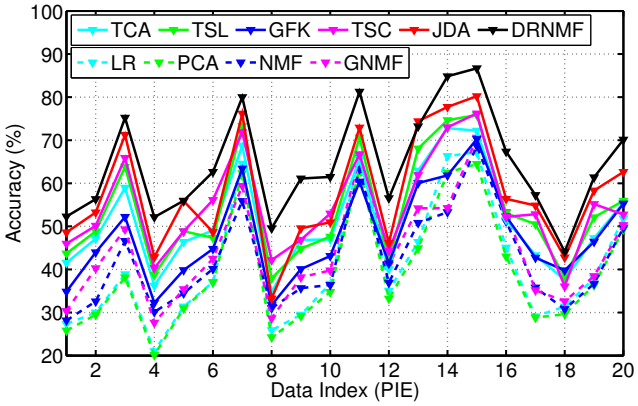


Figure 2: All Results on PIE

JDA, is **6.76**%, implying DRNMF can significantly reduce the classification error by **16.16**%. The results verify the effectiveness and superiority of DRNMF for learning robust visual features for transfer feature learning. In addition, the results also reveal some important points as below.

First, comparing PCA, TCA and JDA to GNMF, TSC and DRNMF respectively, we can observe that GNMF, TSC and DRNMF achieve much better performance. This phenomenon implies that discovering intrinsic hidden semantics (by NMF or SC) of data and considering the geometrical distribution can indeed result in more effective features, and only considering probability distribution is not enough. This is an important reason for the superiority of DRNMF compared to other transfer learning methods which only focus on reducing the distribution difference, which is as well a main motivation for combining NMF to transfer learning.

Second, transfer feature learning methods can markedly outperform standard feature learning methods, since distribution difference is a crucial issue under cross-domain settings which is ignored by standard feature learning methods.

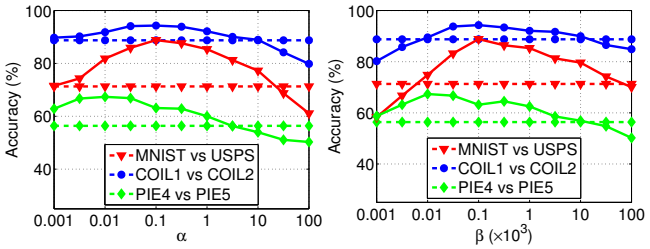
Third, DRNMF and JDA, who take the conditional probability distribution into consideration, outperforms TSC and TCA who ignore it, which validates the importance of conditional probability distribution for transfer feature learning.

Last, DRNMF, which takes all perspectives above, i.e., discovering intrinsic hidden semantics, geometrical distribution, marginal and conditional probability distributions, into consideration, achieves best performance. Combining three points above, we can see that all these perspectives are important and indispensable for learning effective and robust visual features for cross-domain image classification.

5.2.2 *Parameter Sensitivity Analysis*

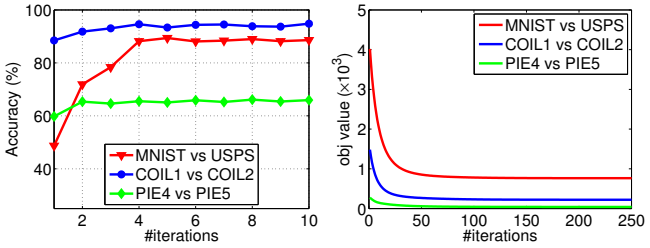
We conduct extensive analysis to validate that DRNMF can achieve stable and superior performance under a wide range of parameter values for α and β, as shown in Figure 3(a) and 3(b) respectively. Because of the limit of space, we just report the results on *MNIST vs USPS*, *COIL1 vs COIL2* and *PIE4 vs PIE5*. Actually, results on other datasets have similar trends. The dashed lines stand for the best baseline results. When we analyze one parameter, we fix the other one to the parameter setting we mentioned in Section 5.1.3.

The parameter α controls the weight of geometrical regularization. If it's too small (e.g., $\alpha < 0.001$), the geometrical distribution can't be preserved in the learned representations which may lower the performance as in [3]. On the



(a) Accuracy w.r.t. α (b) Accuracy w.r.t. β

Figure 3: Parameter sensitivity analysis



(a) Acc. w.r.t. #iteration (b) Obj. w.r.t. #iteration

Figure 4: Other issues

other hand, if it's too large (e.g., $\alpha > 100$), this term may dominate the objective function in Eq. (9) such that the others can't work. And a too large α may also lead to the trivial solution and scale transfer problem [10] as GNMF. We can observe from Figure 3(a) that DRNMF can consistently outperform the best result of baseline methods when $\alpha \in [0.001, 1]$, which is a quite wide range for parameter α.

The weight of probability distribution regularization is controlled by parameter β. If it's too small (e.g., $\beta < 10$), the distribution difference can't be reduced, which may markedly degrade the performance of classifiers under cross-domain setting. On the contrary, if it's too large (e.g., $\beta > 10^5$), only the probability distributions of different domains are matched while the NMF and the geometrical distribution are discarded. As we have proved previously, it's not enough if only the probability distribution is taken into consideration. The result proves again that the intrinsic hidden semantics, geometrical distribution, marginal and conditional probability distributions are all important and indispensable perspectives for TVFL. Fortunately, it's not difficult to choose a proper value for β from $[10, 10^4]$, where DRNMF consistently outperforms the best baseline.

5.2.3 *Other Issues*

Though there is no theoretical evidence that Algorithm 1 can converge w.r.t. outer iteration, the experimental results validates that DRNMF has better performance with more iterations, implying more correct pseudo labels are obtained and conditional probability distributions between domains are better matched indeed with more iterations, which also demonstrates the effectiveness of our EM-like iterative strategy. And superior and stable performance can be reached within 10 outer iterations, as shown in Figure 4(a).

The optimization algorithm in inner iterations (for minimizing Eq. (9)) is theoretically guaranteed to converge as proved in Section 4, but we also care about how fast it can converge. In Figure 4(b), we plot the objective function value (averaged by the number of images) w.r.t. the number

of iterations. We can observe that the objective function value decreases steadily and converges very fast, generally within 100 iterations, which validates the effectiveness of the multiplication updating rules in Eq. (16) and Eq. (17).

6. CONCLUSION

In this paper, we propose a novel method DRNMF for TVFL, who simultaneously uncovers the intrinsic hidden semantics of data, preserves geometrical distribution, and reduces both marginal and conditional probability distribution difference between domains, which are all indispensable and important for TVFL. We carried out extensive experiments on three types of cross-domain image classification tasks, and the results demonstrate that DRNMF can significantly outperform several state-of-the-art related methods, verifying the superiority and effectiveness of DRNMF.

7. ACKNOWLEDGEMENT

This research was supported the National Natural Science Foundation of China (Grant No. 61271394), and the National HeGaoJi Key Project (No. 2013ZX01039-002-002). And the authors would like to thank the anonymous reviewers for their valuable comments.

8. REFERENCES

[1] Y. Aytar and A. Zisserman. Tabula rasa: Model transfer for object category detection. In *International Conference on Computer Vision*, 2011.

[2] L. Bruzzone and M. Marconcini. Domain adaptation problems: A dasvm classification technique and a circular validation strategy. *IEEE TPAMI*, 2010.

[3] D. Cai, X. He, J. Han, and T. S. Huang. Graph regularized nonnegative matrix factorization for data representation. *IEEE TPAMI*, 2011.

[4] M. Chen, K. Q. Weinberger, and J. C. Blitzer. Co-training for domain adaptation. In *Advances in neural information processing systems*, 2011.

[5] Y. Chen, J. Zhang, D. Cai, W. Liu, and X. He. Nonnegative local coordinate factorization for image representation. *IEEE TIP*, 2013.

[6] A. P. Dempster, N. M. Laird, , and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *JRSS*, 39(1):1–38, 1977.

[7] G. Ding, Y. Guo, and J. Zhou. collective matrix factorization hashing for multimodal data. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2014.

[8] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2012.

[9] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Scholkopf, and A. J. Smola. A kernel method for the two-sample problem. In *Advances in neural information processing systems*, 2006.

[10] Q. Gu, C. Ding, and J. Han. On trivial solution and scale transfer problems in graph regularized nmf. In *International Joint Conference on Artificial Intelligence*, 2011.

[11] M. Guillaumin and V. Ferrari. Large-scale knowledge transfer for object localization in imagenet. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2012.

[12] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2006.

[13] I. H. Jhuo, D. Liu, D. T. Lee, and S. Chang. Robust visual domain adaptation with low-rank reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2012.

[14] L. Jie, T. Tommasi, and B. Caputo. Multiclass transfer learning from unconstrained priors. In *International Conference on Computer Vision*, 2011.

[15] C. H. Lampert and O. Krąğomer. Weakly-paired maximum covariance analysis for multimodal dimensionality reduction and transfer learning. In *European Conference on Computer Vision (ECCV)10*.

[16] D. D. Lee and H. S. Seung. Learning the parts of objects by nonnegative matrix factorization. In *Nature*, 1999.

[17] D. D. Lee and H. S. Seung. Learning the parts of objects by nonnegative matrix factorization. In *Advances in neural information processing systems*, 2001.

[18] S. Z. Li, X. Hou, H. Zhang, and Q. Cheng. Learning spatially localized, parts-based representation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2001.

[19] N. K. Logothetis and D. L. Sheinberg. Visual object recognition. *Annual Review of Neuroscience*, 1996.

[20] M. Long, G. Ding, J. Wang, J. Sun, Y. Guo, and P. S. Yu. Transfer sparse coding for robust image representation. In *CVPR*, 2013.

[21] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu. Transfer feature learning with joint distribution adaptation. In *ICCV*, 2013.

[22] S. J. Pan, I. Tsang, J. T. Kwok, and Q. Yang. Transfer learning via dimensionality reduction. In *AAAI*, 2008.

[23] S. J. Pan, I. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *IEEE TNN*, 2011.

[24] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE TKDE*, 2010.

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

[26] B. Quanz, J. Huan, and M. Mishra. Knowledge transfer with low-quality data: A feature extraction issue. *IEEE TKDE*, 2012.

[27] S. D. Roy, T. Mei, W. Zeng, and S. L. Socialtransfer. Cross-domain transfer learning from social streams for media applications. In *ACM MM*, 2012.

[28] S. Si, D. Tao, and B. Geng. Bregman divergence-based regularization for transfer subspace learning. *IEEE TKDE*, 2010.

[29] H. Wang, F. Nie, H. Huang, and C. Ding. Dyadic transfer learning for cross-domain image classification. In *ICCV*, 2011.

[30] E. Zhong, W. Fan, J. Peng, K. Zhang, J. Ren, D. Turaga, and O. Verscheure. Cross domain distribution adaptation via kernel mapping. In *KDD*, 2009.