Robust and Discriminative Self-Taught Learning

Hua Wang HUAWANGCS@GMAIL.COM
Colorado School of Mines, 1610 Illinois Street, Golden, Colorado 80401 USA

Feiping Nie FEIPINGNIE@GMAIL.COM
Heng Huang HENG@UTA.EDU
The University of Texas at Arlington, 500 UTA Boulevard, Arlington, Texas 76019 USA

Abstract

The lack of training data is a common challenge in many machine learning problems, which is often tackled by semi-supervised learning methods or transfer learning methods. The former requires unlabeled images from the same distribution as the labeled ones and the latter leverages labeled images from related homogenous tasks. However, these restrictions often cannot be satisfied. To address this, we propose a novel robust and discriminative self-taught learning approach to utilize any unlabeled data without the above restrictions. Our new approach employs a robust loss function to learn the dictionary, and enforces the structured sparse regularization to automatically select the optimal dictionary basis vectors and incorporate the supervision information contained in the labeled data. We derive an efficient iterative algorithm to solve the optimization problem and rigorously prove its convergence. Promising results in extensive experiments have validated the proposed approach.

1. Introduction

Traditional machine learning methods usually work well when sufficient training data are available. However, because manually labeling data is both expensive and time-consuming, it is desirable to have new techniques to learn a classifier with high accuracy but from only a limited number of labeled training data. Semi-supervised learning methods (Zhu, 2006) exploit unlabeled data to remedy the lack of labeled data, which, however, requires that the unlabeled data are under the same distribution as the labeled. Typical transfer learning methods (Pan & Yang, 2009) relax this restriction to learn useful representations from data under different distributions, which, though, still require the further labeled data from related homogenous tasks. For example, the images from horse, dolphin, bear classes can help categorizing other animal images, such as armadillos, tigers, zebras images. In this paper, we ask how unlabeled data from *heterogeneous* classes, which are much easier to be obtained, to be used for helping classification tasks. For example, given unlimited access to unlabeled and randomly chosen images, *e.g.*, those downloaded from Internet (probably none of which contains the object of interest), can we do better in an existing image categorization task?

Motivated by the observation (Raina et al., 2007; Raina, 2009; Lee et al., 2009) that many randomly downloaded images can still contain the basic visual patterns (such as edges) that are similar to those in existing training images, as shown in Figure 1, one can learn a succinct and higher-level feature representation of the unlabeled data, which could potentially improve the existing image categorizations. Our new approach belongs to an emerging machine learning topic of *self-taught learning (STL)* (Raina et al., 2007; Dai et al., 2008; Raina, 2009; Lee et al., 2009), a special type of transfer learning. Because self-taught learning places fewer restrictions on unlabeled data, it has much more applications than traditional transfer learning or semi-supervised learning methods. For example, it is far easier to obtain 10,000 Internet images than obtain 10,000 images of tigers or armadillos.

The flexibility of self-taught learning makes it of particular use in practice, which, though, also brings new challenges. First, because the unlabeled data are ran-

Proceedings of the 30^{th} International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

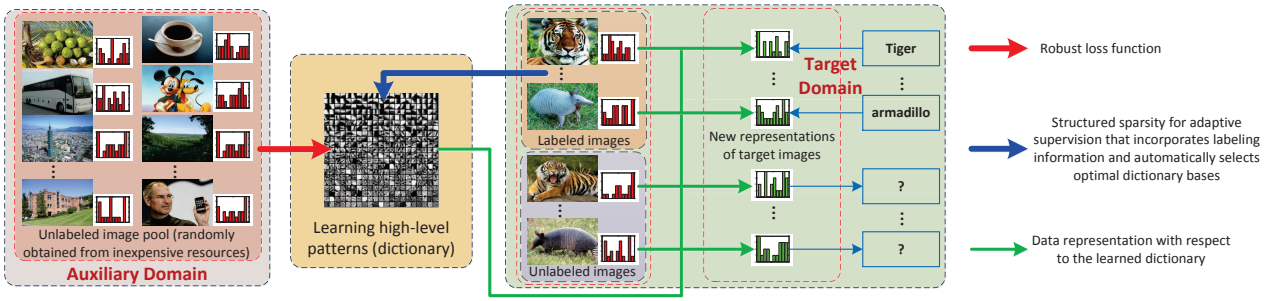


Figure 1. Diagram of the proposed RD-STL method: high-level image patterns (a dictionary) are learned to transfer knowledge from an auxiliary domain with unlimited access of inexpensive images to a target domain in which we are interested to classify images. Different to existing STL methods, (1) we use the robust $\ell_{2,1}$-norm loss function that is insensitive to outliers, which are abundant in the unlabeled auxiliary data due to the nature of STL; (2) we employ the structured sparse regularizations to incorporate the supervision information in the target domain without introducing additional parameter and automatically select the data specific optimal dictionary bases.

domly obtained from Internet or other inexpensive sources, these data could be very different the target objects. Consequently, noises and outlier samples abound in the unlabeled data by nature, which, compared to standard supervised classification tasks, requires more robustness on the learning model. Second, existing self-taught learning methods (Raina et al., 2007; Lee et al., 2009) unsupervisedly learn the feature *dictionary* and ignore the supervision information contained in labeled images. Thus, effectively utilizing the labeling information is another challenging yet important issue. To tackle these difficulties, in this paper we propose a novel robust and discriminative self-taught learning approach with the following contributions:

1. Instead of using traditional squared ℓ_2-norm loss function when learning the feature dictionary, we use the $\ell_{2,1}$-norm loss function, which is robust to outlier samples (Ding et al., 2006; Nie et al., 2010). To our best knowledge, we are the first to learn a robust dictionary in both self-taught learning and dictionary learning areas.

2. Different to existing methods that incorporate prior knowledge by introducing additional terms into the objectives, we propose a new dictionary learning objective to leverage the labeling information by imposing structured sparsity on the representation coefficients via the $\ell_{2,1}$-norm regularizations (Bradley & Bagnell, 2009; Jia et al., 2010), such that no extra parameter is involved and our model is easier to fine tune. Moreover, through the selected prominent basis vectors due to the $\ell_{2,1}$-norm regularization, the optimal dictionary size is automatically determined.

3. We derive a new efficient iterative solution algorithm, whose convergence is rigorously proved.

2. Robust and Discriminative Self-Taught Learning (RD-STL)

In this section, we will first briefly review the traditional self-taught learning method, from which we will systematically develop our new robust and discriminative self-taught learning model.

Notations and problem formalization. Throughout this paper, we will write matrices as bold uppercase characters and vectors as bold lowercase characters. Given a matrix $\mathbf{M} = [m_{ij}]$, we denote its i-th row and its j-th column as $\mathbf{m}^i$ and $\mathbf{m}_j$, respectively.

In self-taught learning, we are given a labeled training set $\left\{\left(\mathbf{x}_i^l, \mathbf{y}_i^l\right)\right\}_{i=1}^n$ (in the target domain as shown in Figure 1) drawn independently and identically from a certain distribution $\mathcal{D}$. $\mathbf{x}_i^l \in \mathbb{R}^p$ is a feature vector associated to its binary label indicator $\mathbf{y}_i^l \in \{0,1\}^K$ for K classes of interest, such that $\mathbf{y}_i^l(k) = 1$ if $\mathbf{x}_i^l$ belongs to the k-th class, and 0 otherwise. In addition, we also have a set of m unlabeled data $\{\mathbf{x}_i^u \in \mathbb{R}^p\}_{i=1}^m$ in the auxiliary domain. Crucially, we do *not* assume that the data from the auxiliary and target domains share the same distribution or class labels. Our goal is to learn from the labeled data $\left\{\mathbf{x}_i^l, \mathbf{y}_i^l\right\}_{i=1}^n$ and the unlabeled data $\{\mathbf{x}_i^u\}_{i=1}^m$ a function that is able to predict labels for an unseen data point $\mathbf{x}$ drawn from the distribution $\mathcal{D}$ in the target domain as shown in Figure 1. For convenience, we write $\mathbf{X}_k \in \mathbb{R}^{p \times n_k}$ $(1 \le k \le K)$ as the data matrix of the k-th class, whose columns are the n_k labeled images belonging to the k-th class. We also write $\mathbf{X}_0 = [\mathbf{x}_1^u, \ldots, \mathbf{x}_m^u]$ for unlabeled images.

2.1. A Brief Review of Traditional STL

In self-taught learning (Raina et al., 2007), we first learn a set of r basis vectors, $\{\mathbf{d}_j \in \mathbb{R}^p\}_{j=1}^r$, forming

a dictionary $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_r] \in \mathbb{R}^{p \times r}$ (allowing $r > p$ to make the dictionary over-complete), from unlabeled data by minimizing the following objective:

$$J_u(\mathbf{D}, \mathbf{a}_i^u) = \sum_{i=1}^{m} \left(\|\mathbf{x}_i^u - \mathbf{D}\mathbf{a}_i^u\|_2^2 + \lambda \|\mathbf{a}_i^u\|_1 \right), \quad s.t.\ \|\mathbf{d}_j\|_2 \leq 1,\ \forall\, 1 \leq j \leq r\ , \tag{1}$$

where $\lambda > 0$ is a parameter and $\mathbf{a}_i^u \in \mathbb{R}^r$ is the representation coefficient vector of $\mathbf{x}_i^u$ with respect to the dictionary $\mathbf{D}$. Here we constrain $\mathbf{d}_j$ to avoid degenerate solution, because the reconstruction errors in the first term of J_u are invariant to simultaneously scaling $\mathbf{D}$ by a scalar and $\mathbf{a}_i^u$ by its inverse (Lee et al., 2007). Due to the ℓ_1-norm regularization on $\mathbf{a}_i^u$, it is sparse with very few non-zero entries (Tibshirani, 1996; Candès & Wakin, 2008). Therefore, $\mathbf{d}_j$ $(1 \leq j \leq r)$ are considered as high-level feature prototypes learned from the unlabeled data, which could be more discriminative and convey more semantic information (Raina et al., 2007; Mairal et al., 2009).

Then the labeled data can be represented with respect to the learned dictionary $\mathbf{D}$ by minimizing:

$$J_l\left(\mathbf{a}_i^l\right) = \left\|\mathbf{x}_i^l - \mathbf{D}\mathbf{a}_i^l\right\|_2^2 + \lambda \left\|\mathbf{a}_i^l\right\|_1,\ \forall\, 1 \leq i \leq n\ , \tag{2}$$

where $\mathbf{a}_i^l \in \mathbb{R}^r$ is the new representation of $\mathbf{x}_i^l$ with respect to $\mathbf{D}$. Again, $\mathbf{a}_i^l$ is sparse due to the ℓ_1-norm regularization.

Finally, the learned $\left\{\mathbf{a}_i^l\right\}_{i=1}^n$ are fed into a classifier, *e.g.*, support vector machine (SVM) as in (Raina et al., 2007), to classify unseen images under the distribution $\mathcal{D}$. Despite a number of imperfections in the current implementations, compared to directly classifying the original feature vectors $\left\{\mathbf{x}_i^l\right\}_{i=1}^n$, self-taught learning have demonstrated better performance in a number of learning tasks (Raina et al., 2007; Raina, 2009; Lee et al., 2009; Bengio et al., 2009). In the rest of this section, we will discuss the weaknesses of the current self-taught learning methods and develop our new method to address them.

2.2. Learning Robust and Adaptively Discriminative Dictionary

Learning robust dictionary with both labeled and unlabeled data. In existing self-taught learning methods (Raina et al., 2007; Raina, 2009; Lee et al., 2009), the dictionary to transfer knowledge is learned from some unlabeled data as in Eq. (1) and used for the labeled data as in Eq. (2), separately. Because the final classification is performed on unseen data under the same distribution as the labeled data, it could be beneficial to learn the dictionary and sparse data representations from both unlabeled and labeled data together by minimizing the following objective:

$$\begin{aligned} \min_{\mathbf{D} \in \mathcal{C}} J_1\left(\mathbf{D}, \mathbf{a}_i^u, \mathbf{a}_i^l\right) = &\sum_{i=1}^{m} \left(\|\mathbf{x}_i^u - \mathbf{D}\mathbf{a}_i^u\|_2^2 + \lambda \|\mathbf{a}_i^u\|_1 \right) \\ &+ \sum_{i=1}^{n} \left(\left\|\mathbf{x}_i^l - \mathbf{D}\mathbf{a}_i^l\right\|_2^2 + \lambda \left\|\mathbf{a}_i^l\right\|_1 \right)\ , \end{aligned} \tag{3}$$

where, for notation brevity, we denote $\mathcal{C} = \left\{\mathbf{D} \middle| \|\mathbf{d}_j\|_2 \leq 1,\ \forall\, 1 \leq j \leq r\right\}$ as the feasible domain of the problem.

Because J_1 in Eq. (3) uses the squared ℓ_2-norm loss function to measure reconstruction errors, which is notoriously known in statistical learning to be sensitive to outlier training samples, following (Ding et al., 2006; Nie et al., 2010) we consider to use a robust loss function to minimize the following objective:

$$\begin{aligned} \min_{\mathbf{D} \in \mathcal{C}} J_2\left(\mathbf{D}, \mathbf{a}_i^u, \mathbf{a}_i^l\right) = &\sum_{i=1}^{m} \left(\|\mathbf{x}_i^u - \mathbf{D}\mathbf{a}_i^u\|_2 + \lambda \|\mathbf{a}_i^u\|_1 \right) \\ &+ \sum_{i=1}^{n} \left(\left\|\mathbf{x}_i^l - \mathbf{D}\mathbf{a}_i^l\right\|_2 + \lambda \left\|\mathbf{a}_i^l\right\|_1 \right)\ . \end{aligned} \tag{4}$$

Denote $\mathbf{X} = [\mathbf{X}_0, \mathbf{X}_1, \ldots, \mathbf{X}_K]$, and $\mathbf{A} = [\mathbf{A}_0, \mathbf{A}_1, \ldots, \mathbf{A}_K]$ where $\mathbf{A}_0 = [\mathbf{a}_1^u, \ldots, \mathbf{a}_m^u]$ and $\mathbf{A}_k = \left[\mathbf{a}_1^l, \ldots, \mathbf{a}_{n_k}^l\right]$, we can write J_2 in a more succinct form using matrices as following:

$$\min_{\mathbf{D} \in \mathcal{C}} J_{\mathrm{R}}(\mathbf{D}, \mathbf{A}) = \left\|(\mathbf{X} - \mathbf{D}\mathbf{A})^T\right\|_{2,1} + \lambda \|\mathbf{A}\|_1\ , \tag{5}$$

Because the reconstruction error terms in Eq. (4) and Eq. (5) are not squared, the outlier samples have less influences and our objectives are more robust.

Learning adaptive dictionary. Because the ℓ_1-norm regularizations used in Eqs. (1–5) flatly enforce sparsity, all the basis vectors, *i.e.*, the underlying data patterns, in the learned dictionary $\mathbf{D}$ are evenly treated and used in the learning process. To capture all potential data patterns, following the theory of compressed sensing (Candès & Wakin, 2008), the dictionary is routinely designed to be over-complete thereby redundant, which makes the subsequent tasks computationally inefficient. Several attempts (Lee et al., 2007; Mairal et al., 2009; 2008; 2010) have been successfully made to address this to learn a compact dictionary with smaller dictionary size. A crucial issue of these methods is that the dictionary size has to be specified by heuristics. The important issue to determine the optimal dictionary size was never taken into account. In this paper, motivated by (Bradley & Bagnell, 2009; Jia et al., 2010) we present a principled method to seek the optimal dictionary basis vectors. Most importantly, following the same idea, the

supervision information can be incorporated with no additional parameter introduced.

Suppose we have an over-complete dictionary $\mathbf{D}$, the basis vector selection can be formalized as:

$$\begin{aligned} \min_{\mathbf{D_X},\mathbf{A}} \ & \left\|(\mathbf{X}-\mathbf{D_X}\mathbf{A})^T\right\|_{2,1} + \lambda\|\mathbf{A}\|_1 , \\ s.t. \ & \mathbf{D_X} \in \mathbf{D},\ |\mathbf{D_X}| = m \ , \end{aligned} \tag{6}$$

where $\mathbf{D_X}$ is the optimal compact dictionary and m is its size. However, three problems impede us to solve Eq. (6) directly. First, $\mathbf{D}$ is not known in a priori, which also needs to be learned. Second, the underlying high-level patterns, *i.e.*, the number of dictionary basis m, are also not known beforehand. Last, Eq. (6) is a combinatorial optimization problem, which is NP-hard. To tackle these difficulties, we first rewrite Eq. (6) in its equivalent form as follows:

$$\min_{\mathbf{D}\in\mathcal{C},\mathbf{A}} \left\|(\mathbf{X}-\mathbf{D}\mathbf{A})^T\right\|_{2,1} + \lambda\|\mathbf{A}\|_1 \ , \ s.t. \ \|\mathbf{A}\|_{2,0} = m \ , \tag{7}$$

where $\|\mathbf{M}\|_{2,0}$ is defined as the number of non-zero rows of the matrix $\mathbf{M}$.

Recent theoretical progresses (Tibshirani, 1996; Candès & Wakin, 2008) show that $\|\mathbf{M}\|_{2,1}$ is the minimum convex hull of $\|\mathbf{M}\|_{2,0}$. When $\mathbf{M}$ is row-sparse enough, one can always minimize $\|\mathbf{M}\|_{2,1}$ to obtain the same result as minimizing $\|\mathbf{M}\|_{2,0}$. Thus, we propose to learn $\mathbf{D}$ and $\mathbf{A}$ from $\mathbf{X}$ by minimizing the following objective:

$$J_{\mathrm{RA}}(\mathbf{D},\mathbf{A}) = \left\|(\mathbf{X}-\mathbf{D}\mathbf{A})^T\right\|_{2,1} + \lambda\|\mathbf{A}\|_{2,1} \ . \tag{8}$$

In Eq. (8), the second term uses the $\ell_{2,1}$-norm regularization, which, different from the flat penalty introduced by the ℓ_1-norm regularization as in Eqs. (5–7), penalizes all n representation coefficients (*i.e.*, all entries in $\mathbf{a}^i$) corresponding to one single basis vector of $\mathbf{D}$ as a whole, and compute the ℓ_1-norm over $\mathbf{a} = \left[\left\|\mathbf{a}^1\right\|_2, \ldots, \left\|\mathbf{a}^r\right\|_2\right]^T$. Consequently, sparsity is conferred on $\mathbf{a}$, and the basis vectors in $\mathbf{D}$ corresponding to the non-zero entries of resulted $\mathbf{a}$ are automatically selected for succeeding data representation. Denote $\mathcal{D}_{\mathbf{X}} = \left\{\mathbf{d}_i \mid \left\|\mathbf{a}^i\right\|_2 > 0\right\}$, *i.e.*, $\mathcal{D}_{\mathbf{X}}$ is a subset of the columns (basis vectors) of $\mathbf{D}$ that correspond to the nonzero entries of $\mathbf{a}$, we may construct $\mathbf{D_X} \in \mathbb{R}^{p\times|\mathcal{D}_{\mathbf{X}}|}$ by using all $\mathbf{d}_i \in \mathcal{D}_{\mathbf{X}}$ as its columns. As a result, $\mathbf{D_X}$ is compact and only the relevant basis vectors specific to the input data are selected, whose number automatically determines the dictionary size. As shown later (in Section 3 of supplementary document due to space limit), we only need to roughly specify a preliminary size of $\mathbf{D}$, which does not impact the dictionary quality of $\mathbf{D_X}$ in a large selection range.

Learning discriminative dictionary. In self-taught learning, because we have both unlabeled and labeled data, we could take advantage of the supervision information in labeled data to make the learned dictionary discriminative thereby benefit the succeeding classifications. Instead of using an additional term to incorporate label information as in most, if not all, prior studies (Mairal et al., 2009; 2008; 2010), we enforce the structured sparsity on the coefficient matrix $\mathbf{A}$ upon the supervision knowledge, such that no extra parameter is required. Specifically, we learn $\mathbf{D}$ and $\mathbf{A}$ from $\mathbf{X}$ by minimizing the following objective:

$$J_{\mathrm{RD}}(\mathbf{D},\mathbf{A}) = \left\|(\mathbf{X}-\mathbf{D}\mathbf{A})^T\right\|_{2,1} + \lambda\sum_{k=0}^{K}\|\mathbf{A}_k\|_{2,1} \ . \tag{9}$$

Upon solution, let $\mathcal{D}_k = \left\{\mathbf{d}_i \mid \left\|\mathbf{a}_k^i\right\|_2 > 0\right\}$ where $\mathbf{a}_k^i$ is the i-th row of $\mathbf{A}_k$, we construct the k-th class specific dictionary $\mathbf{D}_k \in \mathbb{R}^{p\times|\mathcal{D}_k|}$ using all $\mathbf{d}_i \in \mathcal{D}_k$ as its columns. Obviously, $\mathbf{D}_0$ is the dictionary learned for the unlabeled data and $\mathbf{D}_k (1 \le k \le K)$ is discriminatively specific to the data belonging to the k-th class, although all of them are constructed from the globally learned super dictionary $\mathbf{D}$. Again, the size of $\mathbf{D}_k$ is automatically determined by $|\mathcal{D}_k|$.

Because the labeled and unlabeled data come different distributions in self-taught learning, it is reasonable to use $\ell_{2,1}$-norm to group the unlabeled data together in dictionary learning. If our model is applied to semi-supervised learning problems, the ℓ_1-norm should be used between labeled and unlabeled data. From this point, we can also see the differences between self-taught learning and semi-supervised learning.

Finally, we call J_{RD} in Eq. (9) as the proposed Robust and Discriminative Self-Taught Learning (RD-STL) approach, because the learned dictionaries $\mathbf{D}_k (1 \le k \le K)$ ($\mathbf{D}_0$ is not used for classification) are both robust to outlier training samples and adaptively discriminative with respect to the classes in the target domain. Our approaches bridges the unlabeled (auxiliary domain) and labeled (target domain) data and transfers knowledge from the former to the latter as shown in Figure 1, where, crucially, we allow data to come from different distributions.

2.3. Optimization Algorithm and Its Analysis

Because our new objective J_{RD} in Eq. (9) comprises multiple terms of $\ell_{2,1}$-norms, it is difficult to solve in general by existing optimization algorithms. Hence we derive a alternately iterative algorithm to solve the problem[1], which employs the same mechanism of the

[1]The algorithm derivation is supplied in Section 2 of the supplementary document due to space limit.

iteratively re-weighted method (Gorodnitsky & Rao, 1997; Nie et al., 2010; Wang et al., 2011; 2012b;c; 2013) to deal with the non-smooth $\ell_{2,1}$-norm terms.

First, when fixing $\mathbf{A}$, we need to solve the following optimization problem:

$$\min_{\mathbf{D}\in\mathcal{C}} \left\|(\mathbf{X}-\mathbf{D}\mathbf{A})^T\right\|_{2,1} . \tag{10}$$

Then, when we fix $\mathbf{D}$, J_{RD} in Eq. (9) is decoupled into the following subproblems for each $k\,(0\le k\le K)$:

$$\min_{\mathbf{A}_k} \left\|(\mathbf{X}_k-\mathbf{D}\mathbf{A}_k)^T\right\|_{2,1} + \lambda\left\|\mathbf{A}_k\right\|_{2,1} . \tag{11}$$

We alternately solve the problems in Eq. (10) and Eq. (11) to minimize the objective J_{RD} in Eq. (9). Our algorithm is described in Algorithm 1, whose convergence is guaranteed by the follow theorem.

Algorithm 1 An efficient iterative algorithm to solve the objective value of Eq. (9).

Input: $\mathbf{X}\in\mathbb{R}^{p\times\tilde{n}}$.
Output: $\mathbf{D}\in\mathbb{R}^{p\times r}$ and $\mathbf{A}\in\mathbb{R}^{r\times\tilde{n}}$.
1. $t=1$. Initialize diagonal matrices $\mathbf{U}^{(t)}\in\mathbb{R}^{\tilde{n}\times\tilde{n}}$ and $\mathbf{V}_k^{(t)}(0\le k\le K)\in\mathbb{R}^{r\times r}$. Initialize $\mathbf{A}^{(t)}\in\mathbb{R}^{r\times\tilde{n}}$.
while not converge **do**
2. Compute $\hat{\mathbf{X}}=\mathbf{X}(\mathbf{U}^{(t)})^{\frac{1}{2}}$ and $\hat{\mathbf{A}}=\mathbf{A}(\mathbf{U}^{(t)})^{\frac{1}{2}}$. Compute $\mathbf{D}^{(t+1)}=\arg\min_{\mathbf{D}\in\mathcal{C}}\left\|(\hat{\mathbf{X}}-\mathbf{D}\hat{\mathbf{A}})^T\right\|_F^2$.
3. For each $k\,(0\le k\le K)$, compute the i-th column of $\mathbf{A}_k^{(t+1)}$ by $u_{ii}^{(t)}(u_{ii}^{(t)}(\mathbf{D}^{(t+1)})^T\mathbf{D}^{(t+1)}+\mathbf{V}_k^{(t)})^{-1}(\mathbf{D}^{(t+1)})^T(\mathbf{X}_k)_i$, where u_{ii} is the i-th diagonal element of $\mathbf{U}^{(t)}$ and $(\mathbf{X}_k)_i$ is the i-th column of $\mathbf{X}_k$. Construct $\mathbf{A}^{(t+1)}$ by $\mathbf{A}_k^{(t+1)}(1\le k\le K)$.
4. Compute the diagonal matrix $\mathbf{U}^{(t+1)}$, where the i-th diagonal element is $\frac{1}{2\sqrt{\left\|\mathbf{x}_i-\mathbf{D}^{(t+1)}\left(\mathbf{A}^{(t+1)}\right)_i\right\|_2^2+\zeta}}$.
5. For each $k\,(0\le k\le K)$, compute the diagonal matrix $\mathbf{V}_k^{(t+1)}$, where the i-th diagonal element is $(\mathbf{V}_k)_{ii}=\frac{1}{2\sqrt{\left\|\left(\mathbf{A}_k^{(t+1)}\right)^i\right\|_2^2+\xi}}$.
6. $t=t+1$.
end while

Theorem 1 *Algorithm 1 decreases the objective value J_{RD} in Eq. (9) in each iteration till converges.*

Proof: We define $L^{(i)}=\sum_{k=0}^{K}Tr((\mathbf{A}_k^{(i)})^T\mathbf{V}_k^{(t)}\mathbf{A}_k^{(i)})$, where $i=t$ or $t+1$. In each iteration t, according to the Step 2 in the Algorithm 1, we know that

$$\begin{aligned}&Tr((\widetilde{\mathbf{X}}-\mathbf{D}^{(t+1)}\widetilde{\mathbf{A}}^{(t)})\mathbf{U}^{(t)}(\widetilde{\mathbf{X}}-\mathbf{D}^{(t+1)}\widetilde{\mathbf{A}}^{(t)})^T)+\lambda L^{(t)}\\&\le Tr((\widetilde{\mathbf{X}}-\mathbf{D}^{(t)}\widetilde{\mathbf{A}}^{(t)})\mathbf{U}^{(t)}(\widetilde{\mathbf{X}}-\mathbf{D}^{(t)}\widetilde{\mathbf{A}}^{(t)})^T)+\lambda L^{(t)} .\end{aligned} \tag{12}$$

According to Step 3 we know,

$$\begin{aligned}&Tr((\widetilde{\mathbf{X}}-\mathbf{D}^{(t+1)}\widetilde{\mathbf{A}}^{(t+1)})\mathbf{U}^{(t)}(\widetilde{\mathbf{X}}-\mathbf{D}^{(t+1)}\widetilde{\mathbf{A}}^{(t+1)})^T)+\lambda L^{(t+1)}\\&\le Tr((\widetilde{\mathbf{X}}-\mathbf{D}^{(t+1)}\widetilde{\mathbf{A}}^{(t)})\mathbf{U}^{(t)}(\widetilde{\mathbf{X}}-\mathbf{D}^{(t+1)}\widetilde{\mathbf{A}}^{(t)})^T)+\lambda L^{(t)} .\end{aligned} \tag{13}$$

Based on Eq. (12) and Eq. (13), we know

$$\begin{aligned}&Tr((\widetilde{\mathbf{X}}-\mathbf{D}^{(t+1)}\widetilde{\mathbf{A}}^{(t+1)})\mathbf{U}^{(t)}(\widetilde{\mathbf{X}}-\mathbf{D}^{(t+1)}\widetilde{\mathbf{A}}^{(t+1)})^T)+\lambda L^{(t+1)}\\&\le Tr((\widetilde{\mathbf{X}}-\mathbf{D}^{(t)}\widetilde{\mathbf{A}}^{(t)})\mathbf{U}^{(t)}(\widetilde{\mathbf{X}}-\mathbf{D}^{(t)}\widetilde{\mathbf{A}}^{(t)})^T)+\lambda L^{(t)} .\end{aligned} \tag{14}$$

Because it can be verified that (Wang et al., 2012d) for function $f(x)=x-\frac{x^2}{2\alpha}$, given any $x\ne\alpha\in\mathbb{R}$, $f(x)\le f(\alpha)$ holds, together with the definitions of $\mathbf{U}^{(t)}$ and $\mathbf{V}_k^{(t)}(1\le k\le K)$, we can derive that

$$\begin{aligned}&\left\|(\widetilde{\mathbf{X}}-\mathbf{D}^{(t+1)}\widetilde{\mathbf{A}}^{(t+1)})^T\right\|_{2,1}\\&-Tr((\widetilde{\mathbf{X}}-\mathbf{D}^{(t+1)}\widetilde{\mathbf{A}}^{(t+1)})\mathbf{U}^{(t)}(\widetilde{\mathbf{X}}-\mathbf{D}^{(t+1)}\widetilde{\mathbf{A}}^{(t+1)})^T)\le\\&\left\|(\widetilde{\mathbf{X}}-\mathbf{D}^{(t)}\widetilde{\mathbf{A}}^{(t)})^T\right\|_{2,1}\\&-Tr((\widetilde{\mathbf{X}}-\mathbf{D}^{(t)}\widetilde{\mathbf{A}}^{(t)})\mathbf{U}^{(t)}(\widetilde{\mathbf{X}}-\mathbf{D}^{(t)}\widetilde{\mathbf{A}}^{(t)})^T)\end{aligned} \tag{15}$$

and

$$\sum_{k=1}^{K}\left\|\mathbf{A}_k^{(t+1)}\right\|_{2,1}-L^{(t+1)}\le\sum_{k=1}^{K}\left\|\mathbf{A}_k^{(t)}\right\|_{2,1}-L^{(t)} . \tag{16}$$

Adding Eqs. (14)–(16) in the both two sides, we have

$$\begin{aligned}&\left\|(\widetilde{\mathbf{X}}-\mathbf{D}^{(t+1)}\widetilde{\mathbf{A}}^{(t+1)})^T\right\|_{2,1}+\lambda\sum_{k=1}^{K}\left\|\mathbf{A}_k^{(t+1)}\right\|_{2,1}\\&\le\left\|(\widetilde{\mathbf{X}}-\mathbf{D}^{(t)}\widetilde{\mathbf{A}}^{(t)})^T\right\|_{2,1}+\lambda\sum_{k=1}^{K}\left\|\mathbf{A}_k^{(t)}\right\|_{2,1} .\end{aligned} \tag{17}$$

Note that, the equalities in Eqs. (12–17) hold if and only if the objective value converges. Because J_{RD} in Eq. (9) is obviously lower bounded by 0, the objective value of J_{RD} is decreased in each iteration till converges, which completes the proof of Theorem 1. □

2.4. Classification of Test Images

Given a test image $\mathbf{x}$ in target domain and the learned dictionaries $\mathbf{D}_k$ $(1\le k\le K)$, we can compute the sparse representation of $\mathbf{x}$ for the k-th class, $\mathbf{a}^{(k)}$, by solving $\min_{\mathbf{a}^{(k)}}\left\|\mathbf{x}-\mathbf{D}_k\mathbf{a}^{(k)}\right\|_2^2+\lambda\left\|\mathbf{a}^{(k)}\right\|_1$. Thus the reconstruction error of $\mathbf{x}$ with respect to the k-th class is computed as $\mathbf{e}^{(k)}=\left\|\mathbf{x}-\mathbf{D}_k\mathbf{a}^{(k)}\right\|_2$. Following the same way, we can compute the reconstruction errors $\mathbf{e}_i^{(k)}$ for labeled images. Then we can compute the adaptive decision boundary (Wang et al., 2009; 2012a) to classify the test image, which can be applied to both single-label and multi-label data sets.

3. Related Methods

Transfer learning and self-taught learning. From machine learning perspective of view, our approach belongs to the important topic of transfer learning, which aims to make use of knowledge, either unsupervised or supervised, from another domain with different distribution to improve the learning in the current domain of interest. We refer readers to (Pan & Yang, 2009) for a comprehensive survey.

Self-taught learning is an emerging branch of transfer learning, which was first formalized in (Raina et al., 2007) and further developed in (Dai et al., 2008; Raina, 2009; Lee et al., 2009). Self-taught learning aims to utilize unlabeled data with as minimum restrictions as possible. The proposed approach is motivated by and closely related to (Raina et al., 2007), yet different from it in a number of important aspects as detailed in Section 2.2, including (1) joint data utilization, (2) robustness to outliers samples that abound in unlabeled images by nature, (3) dictionary discriminativity and (4) optimal dictionary size selection.

Sparse coding and dictionary learning. Sparsity is one of the intrinsic properties of real world data (Tibshirani, 1996), which makes it useful in many machine learning and computer vision tasks, such as face recognition (Mairal et al., 2009), image classification (Bengio et al., 2009), digital art authentication (Mairal et al., 2010), and many others.

Recent studies (Lee et al., 2007) have demonstrated that decomposing a signal using a few atoms of learned dictionary often leads to state-of-the-art results in real world applications, which aroused considerable interest in the machine vision community (Lee et al., 2007; Bengio et al., 2009; Mairal et al., 2010; 2009). Although a variety of aspects of dictionary learning have been addressed by these previous works, none of them takes into account the dictionary robustness problem. Moreover, these methods typically pre-specify the dictionary size heuristically or by prior knowledge, while how to determine the optimal dictionary size in a principled way is much less studied (Bradley & Bagnell, 2009; Jia et al., 2010). In addition, existing supervised dictionary learning methods routinely employ an additional term to incorporate the labeling information, which inevitably complicates the learning models and makes them less practically useful. Contrastly, our new RD-STL approach gracefully solves all these important yet challenging problems in a unified framework via joint $\ell_{2,1}$-norm minimizations, which makes our model of particular use in real-world applications.

4. Experimental Results

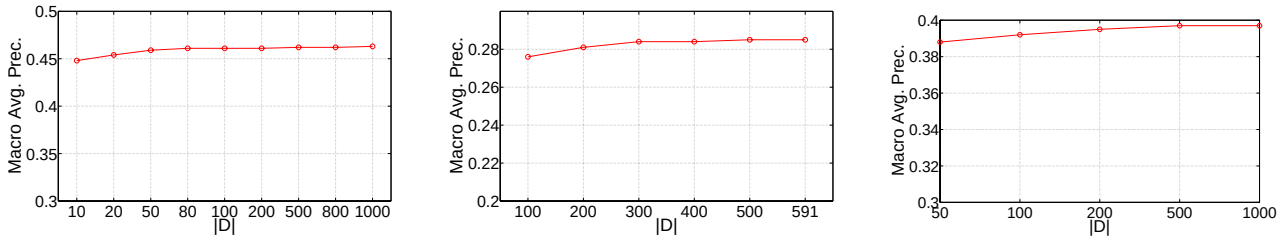
In this section, we experimentally evaluate the proposed approach, where our goal is to examine its capability to improve the classification performance in the target domain by taking advantage of unlabeled data that come from an inexpensive source with distributions (possibly) different from the target data.

Unlabeled images in auxiliary domain. We randomly downloaded 5000 images from the **LabelMe** data set and use them as unlabeled images in the auxiliary domain. Because the more than 10 thousands images in the LabelMe data set come from numerous resources, including Internet, video clips, daily photo, *etc*., it is an ideal source for unlabeled images in self-taught learning.

Labeled and test images in target domain. We use **TRECVID 2005**, **MSRC-v2** and **NUS-WIDE-Object** images data sets as target data sets, which are broadly used in computer vision studies. The details of the data sets are supplied in Section 1 of the supplementary document due to space limit. Our goal is to classify the test images in these data sets using the proposed RD-STL model.

4.1. Study of the Size of the Preliminary Dictionary $|\mathbf{D}|$

In the proposed method, a preliminary dictionary $\mathbf{D}$ is learned, from which we adaptively select discriminative basis vectors specific for each class. Therefore, although the sizes of the ultimate output dictionaries $\mathbf{D}_k\,(0 \leq k \leq K)$ are automatically determined by the learned patterns in $\mathbf{A}_k\,(0 \leq k \leq K)$, we still need to pre-specify the size of $\mathbf{D}$, same as in existing related works (Raina et al., 2007; Raina, 2009; Bengio et al., 2009; Lee et al., 2007; Mairal et al., 2009; 2008; Zhang & Li, 2010; Pham & Venkatesh, 2008; Mairal et al., 2010). However, different from these prior studies that directly use $\mathbf{D}$ for data representation and classification, the qualities, *i.e*., the subsequence classification accuracies, of learned adaptive dictionaries $\mathbf{D}_k\,(0 \leq k \leq K)$ do not heavily rely on the size of $\mathbf{D}$. As shown in Figure 2, when the sizes of the preliminary dictionaries vary in a large range, the classification accuracies of the proposed method remain considerably stable. This demonstrates that our new method is able to adaptively learn class specific dictionary for classification, which do no rely on the preliminary dictionary size as long as it is not too small. Empirically, when $|\mathbf{D}| \geq 2K$, the subsequent classification accuracy is generally satisfactory. In all our experiments, we set $|\mathbf{D}| = \min\{1000, n\}$.



(a) TRECVID 2005. (b) MSRC-v2. (c) NUS-WIDE-Object.

Figure 2. Classification accuracies of the proposed RD-STL method on the three experimental data sets with respect to the size of the preliminary dictionary $|\mathbf{D}|$.

4.2. Improved Image Categorization

We compare our approach to the following related methods: supervised learning method (1) SVM as baseline; two widely used semi-supervised learning methods including (2) transductive SVM (TSVM) (Joachims, 1999) method and (3) the Green's function (GF) method (Ding et al., 2007); and three transfer learning methods including (4) knowledge transfer by words (KTW) method (Li et al., 2009), (5) self-taught clustering (STC) method (Dai et al., 2008) and (6) self-taught learning (STL) method (Raina et al., 2007). STL method needs to use SVM to classify the learned target data representations. We implement these compared methods and fine tune their parameters following their original works. For SVM and TSVM methods, we use the Gaussian kernel (*i.e.*, $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2\right)$), and fine tune γ and the regularization parameter C in the range of $\left\{10^{-5}, \ldots, 10^{-1}, 1, 10^1, \ldots, 10^5\right\}$. Because these methods are single-label classification methods while the target image data sets are multi-label data sets, following (Wang et al., 2010a) we employ the one-*vs*.-other strategy to deal with multi-label data.

In addition, we also compare our approach against two most recent multi-label classification methods including (7) multi-label feature transform (MLFT) (Wang et al., 2010b) method and (8) Multi-Label Least Square (MLLS) (Ji et al., 2010) method.

We implement five versions of the proposed approach to evaluate the usefulness of its component terms as following: (A) the simplest joint self-taught learning method using J_1 in Eq. (3), denoted as "J-STL"; (B) robust self-taught learning method using J_{R} in Eq. (5), denoted as "R-STL"; (C) robust and adaptive self-taught learning method using J_{RA} in Eq. (8), denoted as "RA-STL"; (D) Discriminative self-taught learning method but not taking into account robustness against outlier samples that minimizes $J_{\mathrm{D}}(\mathbf{D}, \mathbf{A}) = \underset{\mathbf{D}\in\mathcal{C}}{}$ $\left\|(\mathbf{X} - \mathbf{D}\mathbf{A})^T\right\|_{\mathrm{F}}^2 + \lambda \sum_{k=0}^{K} \left\|\mathbf{A}_k\right\|_{2,1}$, denoted as "D-STL"; and (E) the proposed RD-STL approach using J_{RD} in Eq. (9). Note that, the first three methods are unsupervised learning methods, therefore we employ SVM to classify the learned target data representations, same as the STL method (Raina et al., 2007). For the last two methods, we classify unseen images using the rules introduced in Section 2.4. For all the five methods, we fine tune the parameter λ by searching the grid of $\left\{10^{-5}, \ldots, 10^{-1}, 1, 10^1, \ldots, 10^5\right\}$.

Performance metrics. Because we experiment with multi-label data sets, following (Wang et al., 2009; 2010a;b), we use the four standard multi-label classification performance metrics, macro averaged precision and F1 score, and micro averaged precision and F1 score, to evaluate the compared methods.

Experimental results. We perform standard 5-fold cross-validation to evaluate the compared methods on the three target data sets. An internal 5-fold cross-validation is conducted in the training data of each of the 5 trials to fine tune the parameters of the compared methods. The experimental results are reported in Table 1, from which we have a number of interesting observations as following.

First, the proposed methods are consistently better than other compared methods, which demonstrate their effectiveness in the task of automatic image categorization.

Second, SVM, TSVM and GF methods do not have satisfactory classification performance. This can be explained as follows. SVM is a supervised method, which can be learned only from the target data while the large amount of auxiliary data are not used. Although the two semi-supervised methods employ both axuliary and target data, they assume them to come from a same distribution, which, however, is not true in both this experiment and many real world applications. That is, a simple mixture of the auxiliary and target data can not leads to satisfactory classification performance.

Third, the two transfer learning methods, KTW and

Table 1. Classification results of compared methods on the three multi-label image data sets.

Methods	TRECVID 2005				MSRC-v2				NUS-WIDE-Object			
	Macro avg.		Micro avg.		Macro avg.		Micro avg.		Macro avg.		Micro avg.	
	Prec.	F1	Prec.	F1	Prec.	F1	Prec.	F1	Prec.	F1	Prec.	F1
SVM	0.269	0.236	0.252	0.291	0.247	0.275	0.234	0.293	0.312	0.314	0.337	0.332
TSVM	0.315	0.298	0.304	0.322	0.253	0.286	0.252	0.303	0.329	0.321	0.349	0.361
GF	0.108	0.151	0.107	0.167	0.121	0.144	0.130	0.161	0.301	0.311	0.320	0.322
MLFT	0.421	0.398	0.420	**0.521**	0.256	0.304	0.259	0.312	0.360	0.409	0.411	0.426
MLLS	0.272	0.275	0.279	0.295	0.255	0.291	0.256	0.301	0.359	0.406	0.404	0.420
KTW	0.403	0.280	0.272	0.273	0.231	0.286	0.244	0.294	0.334	0.376	0.381	0.401
STC	0.404	0.281	0.270	0.275	0.236	0.284	0.247	0.289	0.342	0.381	0.387	0.411
STL	0.425	0.315	0.357	0.296	0.261	0.306	0.267	0.311	0.365	0.413	0.417	0.435
J-STL	0.426	0.320	0.359	0.303	0.265	0.311	0.271	0.316	0.369	0.417	0.420	0.438
R-STL	0.451	0.353	0.361	0.337	0.276	0.326	0.287	0.323	0.381	0.436	0.434	0.451
RA-STL	0.452	0.375	0.389	0.461	0.279	0.337	0.295	0.327	0.383	0.435	0.437	0.450
D-STL	0.438	0.336	0.361	0.423	0.270	0.319	0.276	0.321	0.377	0.430	0.429	0.441
RD-STL	**0.463**	**0.411**	**0.442**	0.520	**0.285**	**0.351**	**0.304**	**0.334**	**0.397**	**0.451**	**0.449**	**0.462**

STC, also do not work well in the experiments. This is because typical transfer learning methods, such as the two used in our experiments, require labeled images in auxiliary data and aim to transfer such knowledge. When prior knowledge are not available in auxiliary data, these two methods actually perform unsupervised clustering on the target data, though with the aid from the auxiliary data. These results clearly show the necessity of STL.

Fourth, compared to STL method, joint self-taught learning by J-STL method does not essentially improve the classification accuracy, which can be seen by the fact that the target data are much less than the unlabeled auxiliary data.

Five, R-STL, RA-STL and RD-STL show much better results than other compared methods, which demonstrate that enhancing robustness against the noises and outliers in the unlabeled source data does improve the classification performance significantly, as expected. In addition, although labeling information is used in D-STL method, its performance is inferior due to not taking into account robustness. Finally, robustness plus discriminativity, *i.e.*, the propose RD-STL approach, achieves the best classification performances, which concretely confirm that these two issues are the most important impact factors to the classification performance of self-taught learning. (See more experimental results and discussion on the robustness of the proposed RD-STL model in Section 4 of the supplementary document due to space limit.)

Last, but not least, although RA-STL method does not outperform R-STL method very much in terms of classification accuracy, the dictionary size of the former is much smaller than that of the latter. The sizes of the learned dictionary $\mathbf{D_X}$ for the three data sets are 46, 29 and 26 respectively, which are much smaller than the preliminary dictionary size for $\mathbf{D}$ as 1000, 591, and 1000 respectively. The same observation also applies to the proposed RD-STL approach. We explain this observation as follows. In traditional sparse learning, motivated by compressed sensing, dictionaries are generally designed to be over-complete. However, the number of underlying patterns of most real-world data is usually small. From information theory perspective, many basis vectors in the learned dictionary are indeed redundant, which might be detrimental to the subsequent sparse solver. To address this, our new method uses data adaptation via the $\ell_{2,1}$-norm regularization to find out the most representative dictionary basis vectors, which leads to a small number of significant dictionary bases and reduces the computational loads in subsequent data representations for test data.

5. Conclusions

To tackle the difficulty of lacking training data in real-world applications, we proposed a novel RD-STL approach to leverage unlabeled images. Different from traditional semi-supervised learning and transfer learning methods, our new approach places significantly fewer restrictions on the unlabeled data. We addressed two important issues in existing self-taught learning methods, including the robustness against noisy and outlier samples in unlabeled data and the usage of supervision information in the target data, by a joint $\ell_{2,1}$-norms minimization framework. Promising results of extensive empirical studies demonstrated the effectiveness of the proposed approach.

Acknowledgments

Corresponding Author: Heng Huang (heng@uta.edu)
This work was partially supported by NSF CCF-0830780, CCF-0917274, DMS-0915228, IIS-1117965.

References

Bengio, S., Pereira, F., Singer, Y., and Strelow, D. Group sparse coding. In *NIPS*, 2009.

Bradley, D.M. and Bagnell, J.A. Convex coding. In *UAI*, 2009.

Candès, E. and Wakin, M. An introduction to compressive sensing. *IEEE Signal Processing Magazine*, 25(2):21–30, 2008.

Dai, W., Yang, Q., Xue, G.R., and Yu, Y. Self-taught clustering. In *ICML*, 2008.

Ding, C., Zhou, D., He, X., and Zha, H. R1-PCA: rotational invariant L 1-norm principal component analysis for robust subspace factorization. In *ICML*, 2006.

Ding, C., Simon, H.D., Jin, R., and Li, T. A learning framework using Green's function and kernel regularization with application to recommender system. In *SIGKDD*, 2007.

Gorodnitsky, I.F. and Rao, B.D. Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm. *IEEE Transactions on Signal Processing*, 45(3):600–616, 1997.

Ji, S., Tang, L., Yu, S., and Ye, J. A shared-subspace learning framework for multi-label classification. *ACM TKDD*, 4(2):1–29, 2010.

Jia, Y., Salzmann, M., and Darrell, T. Factorized latent spaces with structured sparsity. In *NIPS*, 2010.

Joachims, T. Transductive inference for text classification using support vector machines. In *ICML*, 1999.

Lee, H., Battle, A., Raina, R., and Ng, A.Y. Efficient sparse coding algorithms. In *NIPS*, 2007.

Lee, H., Raina, R., Teichman, A., and Ng, A.Y. Exponential family sparse coding with applications to self-taught learning. In *IJCAI*, 2009.

Li, T., Sindhwani, V., Ding, C., and Zhang, Y. Knowledge transformation for cross-domain sentiment classification. In *SIGIR*, 2009.

Mairal, J., Bach, F., Ponce, J., Sapiro, G., and Zisserman, A. Discriminative learned dictionaries for local image analysis. In *CVPR*, pp. 1–8, 2008.

Mairal, J., Bach, F., and Ponce, J. Task-Driven Dictionary Learning. *Tech Report, INRIA*, 2010.

Mairal, Julien, Bach, Francis, Ponce, Jean, Sapiro, Guillermo, and Zisserman, Andrew. Supervised dictionary learning. In *NIPS*, 2009.

Nie, F., Huang, H., Cai, X., and Ding, C. Efficient and Robust Feature Selection via Joint l2,1-Norms Minimization. In *NIPS*, 2010.

Pan, S.J. and Yang, Q. A survey on transfer learning. *IEEE TKDE*, 2009. ISSN 1041-4347.

Pham, D.S. and Venkatesh, S. Joint learning and dictionary construction for pattern recognition. In *CVPR*, pp. 1–8, 2008.

Raina, R. Self-taught learning. *PhD thesis of Stanford University*, 2009.

Raina, R., Battle, A., Lee, H., Packer, B., and Ng, A.Y. Self-taught learning: Transfer learning from unlabeled data. In *ICML*, 2007.

Tibshirani, R. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

Wang, H., Huang, H., and Ding, C. Image annotation using multi-label correlated Green's function. In *ICCV*, 2009.

Wang, H., Ding, C., and Huang, H. Multi-Label Classification: Inconsistency and Class Balanced K-Nearest Neighbor. In *AAAI*, 2010a.

Wang, H., Huang, H., and Ding, C. Multi-label Feature Transform for Image Classifications. *ECCV*, 2010b.

Wang, Hua, Nie, Feiping, Huang, Heng, Risacher, Shannon, Ding, Chris, Saykin, Andrew J, and Shen, Li. Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance. In *ICCV 2011*, pp. 557–562. IEEE, 2011.

Wang, Hua, Huang, Heng, and Ding, Chris. Function-function correlated multi-label protein function prediction over interaction networks. In *Research in Computational Molecular Biology (RECOMB 2012)*, pp. 302–313. Springer, 2012a.

Wang, Hua, Nie, Feiping, Huang, Heng, Kim, Sungeun, Nho, Kwangsik, Risacher, Shannon L, Saykin, Andrew J, Shen, Li, et al. Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the adni cohort. *Bioinformatics*, 28(2):229–237, 2012b.

Wang, Hua, Nie, Feiping, Huang, Heng, Risacher, Shannon L, Saykin, Andrew J, Shen, Li, et al. Identifying disease sensitive and quantitative trait-relevant biomarkers from multidimensional heterogeneous imaging genetics data via sparse multimodal multitask learning. *Bioinformatics*, 28(12):i127–i136, 2012c.

Wang, Hua, Nie, Feiping, Huang, Heng, Yan, Jingwen, Kim, Sungeun, Risacher, Shannon, Saykin, Andrew, and Shen, Li. High-order multi-task feature learning to identify longitudinal phenotypic markers for alzheimer's disease progression prediction. In *NIPS*, 2012d.

Wang, Hua, Nie, Feiping, Huang, Heng, and Ding, Chris. Heterogeneous Visual Features Fusion via Sparse Multimodal Machine. In *CVPR 2013*, 2013.

Zhang, Q. and Li, B. Discriminative K-SVD for dictionary learning in face recognition. In *CVPR*, pp. 2691–2698, 2010.

Zhu, X. Semi-supervised learning literature survey. *Technical report, University of Wisconsin-Madison*, 2006.