Discriminative Semi-Supervised Dictionary Learning with Entropy Regularization for Pattern Classification

Meng Yang,[1 2] Lin Chen[1]
[1]College of Computer Science & Software Engineering, Shenzhen University, Shenzhen, China
[2]School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China

Abstract

Dictionary learning has played an important role in the success of sparse representation, which triggers the rapid developments of unsupervised and supervised dictionary learning methods. However, in most practical applications, there are usually quite limited labeled training samples while it is relatively easy to acquire abundant unlabeled training samples. Thus semi-supervised dictionary learning that aims to effectively explore the discrimination of unlabeled training data has attracted much attention of researchers. Although various regularizations have been introduced in the prevailing semi-supervised dictionary learning, how to design an effective unified model of dictionary learning and unlabeled-data class estimating and how to well explore the discrimination in the labeled and unlabeled data are still open. In this paper, we propose a novel discriminative semi-supervised dictionary learning model (DSSDL) by introducing discriminative representation, an identical coding of unlabeled data to the coding of testing data final classification, and an entropy regularization term. The coding strategy of unlabeled data can not only avoid the affect of its incorrect class estimation, but also make the learned discrimination be well exploited in the final classification. The introduced regularization of entropy can avoid overemphasizing on some uncertain estimated classes for unlabeled samples. Apart from the enhanced discrimination in the learned dictionary by the discriminative representation, an extended dictionary is used to mainly explore the discrimination embedded in the unlabeled data. Extensive experiments on face recognition, digit recognition and texture classification show the effectiveness of the proposed method.

Introduction

Inspired by the sparsity mechanism of human vision system (Olshausen and Field 1996), sparse representation has been widely applied to many tasks, such as image processing (Aharon, Michael, and Alfred 2006; Elad and Aharon 2006) and image classification (Mairal et al. 2008; Wright et al. 2009; Wagner et al. 2010; Yang et al. 2009). A basic model of sparse presentation can be roughly written as $\boldsymbol{y} \approx \boldsymbol{Dx}$, which $\boldsymbol{D}$ is a dictionary and $\boldsymbol{x}$ is a sparse coding vector. The dictionary $\boldsymbol{D}$, which should faithfully represent an input signal regardless of various variations, plays an important role in the success of sparse representation (Rubinstein, Bruckstein and Elad 2010). Compared to the handcrafted and off-the-shelf representation bases, dictionary learning has been reported leading performance in image denoising (Bryt and Elad 2008; Zhou et al. 2010), image compression (Bryt and Elad 2008), face recognition (Wright et al. 2009; Zhang and Li 2010), and image classification (Huang and Aviyente 2006; Mairal et al. 2009; Pham and Svetha 2008; Yang and Zhang 2014; Mairal, Bach and Ponce 2012).

Unsupervised dictionary learning has been developed and widely applied into image denoising (Bryt and Elad 2008; Zhou et al. 2010) and image compression (Bryt and Elad 2008). The desired dictionary is expected to only have a powerful representation ability since no class information is used in the unsupervised dictionary learning. One representative unsupervised dictionary learning (DL) is the KSVD algorithm (Aharon, Michael and Alfred 2006), which has achieved promising abilities in image compression and image restoration by learning an over-complete dictionary for local image patches.

Due to the lack of label information, unsupervised dictionary learning is powerful for data reconstruction, but not advantageous for classification tasks. In the tasks of image classification with labeled training data provided, how to effectively explore the class information and design supervised discriminative dictionary becomes a hot topic in the fields of image classification and sparse representation.

Supervised dictionary learning methods has been extensively studied in recent years for the case that enough labeled training samples are available. As reviewed in Section 2, supervised dictionary learning models usually explore the discrimination of dictionary by requiring the powerful classification ability of coding coefficients, the dis-

criminative representation ability of class-specific dictionary, or both. Most of supervised dictionary learning methods can achieve nice performance when the labeled training samples are sufficient and training images have a good quality. However, the labeled training data is expensive and difficult to obtain due to the vast human effort involved. On the other hand, there are abundant unlabeled images that can be collected easily from public (*e.g.*, for digit recognition huge samples can be collected from public). The above status has motivated many researchers to develop semi-supervised learning.

There are mainly four popular categories in semi-supervised learning: Co-Training (Blum and Mitchell 1998), graph-based semi-supervised learning (Zhu 2005), semi-supervised support vector machines (S3VM) (Sindhwani and Keerthi 2006), and semi-supervised dictionary learning. Due to impressive performance of sparse representation and dictionary learning, a variety of semi-supervised dictionary learning methods (Pham and Svetha 2008; Zhang, Jiang and Davis 2012; Shrivastava et al. 2012; Mohamadabadi 2013; Wang et al 2013; Wang, Guo and Li 2015) have been proposed. Based on the relationship between dictionary atoms and class labels, the prevailing semi-supervised dictionary learning can be divided into two categories: semi-supervised class-specific dictionary learning and semi-supervised shared dictionary learning.

Inspired by Fisher discrimination dictionary learning (Yang and Zhang 2011), Shrivastava et al. (Shrivastava et al. 2012) learnt a class-specific semi-supervised dictionary. By manually designing a class possibility estimation function for unlabeled data, a set of class-specific dictionaries are learned from labeled and unlabeled training data. However, its dictionary learning model is a little complex and the class probability of unlabeled training samples is artificially designed but not derived from the objective function.

Most of semi-supervised dictionary learning methods aim to learn a shared dictionary. Pham et al. (Pham and Svetha 2008) incorporated the reconstruction error of both the labeled and unlabeled data with sparsity constraint into a joint objective function. Zhang et al. (Zhang, Jiang and Davis 2012) proposed an online semi-supervised dictionary learning model, in which the reconstruction error of both labeled data and unlabeled data, label consistency and the classification error were integrated into a joint model. In these two semi-supervised dictionary methods mentioned above, the unlabeled training data is only used to learn a shared dictionary, ignoring to explore the discrimination hidden in the unlabeled data.

Wang et al.(Wang et al. 2013) proposed a robust dictionary learning method by utilizing a new $l_{2,0+}$-norm loss function to measure the reconstruction errors and exploiting the structural sparse regularization of labeled and unlabeled data. Recently, Wang et.al (Wang, Guo and Li 2015) proposed an adaptively unified semi-supervised dictionary learning model which integrated the reconstruction error of both the labeled data and unlabeled data, and classifier learning into a unified framework. Meanwhile, the weights of unlabeled samples will be assigned adaptively by a penalty function. Wang et al. (Wang et al. 2016) grouped the unlabeled samples by using the coefficient-based relationship between the labeled and unlabeled samples. Although the classifier based on the coding coefficient associated to the shared dictionary is adopted, the powerful class-specific representation ability cannot be used.

Although several semi-supervised dictionary learning approaches have been proposed. How to effectively utilize the discrimination of unlabeled data and build a unified model of class-specific dictionary learning and class possibility estimating of unlabeled data are still open. In this paper, we propose a novel method to learn discriminative dictionaries in semi-supervised manner by introducing discriminative representation of training data, an identical coding of unlabeled data to the coding of testing data in final classification, and an entropy regularization term. The coding strategy of unlabeled data can not only avoid the affect of incorrect class estimation, but also make the discrimination of the learned dictionary be well exploited in the final classification. The introduced regularization of entropy can avoid overemphasizing on some uncertain estimated classes for unlabeled samples. Apart from the enhanced discrimination in the learned dictionary by the discriminative coding of labeled data, an extended dictionary is used to mainly explore the discrimination embedded in the unlabeled data. Experimental results on face recognition, digit recognition and texture classification clearly demonstrate that our algorithm can achieve superior performance in the semi-supervised classification tasks.

Related Work

Supervised dictionary learning

For supervised shared dictionary learning, discrimination of the universal dictionary was ordinarily explored by jointly learning a dictionary and a classifier over the coding coefficients (Zhang and Li 2010; Mairal et al. 2009; Mairal, Bach and Ponce 2012; Jiang, Lin and Davis 2013). With the learnt universal dictionary, the generated coding coefficient, which is expected to be discriminative, is used to conduct classification. For instance, Zhang and Li (Zhang and Li 2010) proposed a joint learning algorithm called discriminative KSVD (DKSVD) for face recognition, followed by the work proposed by Jiang *et al.* (Jiang, Lin and Davis 2013) via adding a label consistent term.

Due to the promising performance of class-specific dictionary representation reported in (Wright et al. 2009), how to learn a structured dictionary, in which each dictionary

atom has a unique class label, have been a hot topic in the field of dictionary learning. Regularizations for class-specific dictionary, *e.g.*, low class-particular dictionary coherence (Ramirez, Sprechmann and Sapiro 2010), good class-specific representation for some class but bad for all the other classes (Mairal et al. 2008; Castrodad and Sapiro 2012), and Fisher discrimination on class-specific dictionary and coding coefficient (Yang and Zhang 2014) has been introduced in the phase of dictionary learning.

In order to utilize the powerful class-specific representation ability and reduce the correlation possibly existing in different class-specific dictionary, hybrid dictionary (Kong and Wang 2012; Zhou and Fan 2012; Yang et al.2014), including universal dictionary atoms and class-specific dictionary atoms, was also proposed. Due to the powerful discrimination of class-specific dictionary shown in previous hybrid dictionary learning work, we proposed to discriminatively learn semi-supervised class-specific dictionaries, without considering the universal dictionary atoms to reduce the complexity of dictionary learning model.

Semi-supervised learning

In other popular semi-supervised learning, Co-Training (Blum and Mitchell 1998) utilizes the multiple views of each sample and selects confident unlabeled samples in one classifier to update the other classifier. A representative graph-based supervised learning method is label propagation (LP) (Zhu 2005), which has been widely used in image classification. S3VM (Sindhwani and Keerthi 2006) attempts to regulate over original SVM framework and adjust decision boundaries by exploring unlabeled data.

Discriminative semi-supervised dictionary learning

To solve the issues, such as how to utilize the power of class-specific dictionary representation and how to estimate the class possibility of unlabeled data, we propose a discriminative semi-supervised dictionary learning (DSSDL) with entropy regularization.

Model of DSSDL with entropy regularization

Let $\boldsymbol{A}=[\boldsymbol{A}_1,...,\boldsymbol{A}_i,...,\boldsymbol{A}_C,\boldsymbol{B}]$ be the training data, where $\boldsymbol{A}_i$ is the i^{th}-class training data and each column of $\boldsymbol{A}_i$ is a training sample, and $\boldsymbol{B}=[\boldsymbol{b}_1,...,\boldsymbol{b}_j,...,\boldsymbol{b}_N]$ is the N training samples with unknown labels from 1 to C. For the semi-supervised learning case, the training samples without labels may not belong to the C classes. However, as many prevailing semi-supervised methods (Pham and Svetha 2008; Zhang, Jiang and Davis 2012; Shrivastava et al. 2012; Mohamadabadi 2013; Wang, Guo and Li 2015) we focus on the case that the identity of unlabeled training data is between 1 and C.

Different from supervised dictionary learning, we divide the desired dictionary into two parts $\boldsymbol{D}=[\boldsymbol{D}_1,...,\boldsymbol{D}_i,...,\boldsymbol{D}_C]$ and $\boldsymbol{E}=[\boldsymbol{E}_1,...,\boldsymbol{E}_i,...,\boldsymbol{E}_C]$, where $\boldsymbol{D}_i$ is an i^{th}-class supervised dictionary that can be initialized with $\boldsymbol{A}_i$, while $\boldsymbol{E}_i$ is an i^{th}-class extended dictionary that mainly explore the discrimination of unlabeled training data. Both $\boldsymbol{D}_i$ and $\boldsymbol{E}_i$ are associated to class i, and they are required to well represent i^{th}-class data but with a bad representation ability for all the other classes.

Let $P_{i,j}$ indicate the relationship between the j^{th} unlabeled training samples and i^{th} class. Then the proposed discriminative semi-supervised dictionary learning (DSSDL) with entropy regularization is formulated as

$$\begin{aligned}\min_{\boldsymbol{D},\boldsymbol{E},\boldsymbol{P},\boldsymbol{X}} &\sum_{i=1}^{C}\left(\|\boldsymbol{A}_i-[\boldsymbol{D}_i\ \boldsymbol{E}_i]\boldsymbol{X}_i^i\|_F^2+\gamma\|\boldsymbol{X}_i^i\|_1+\lambda\|\boldsymbol{X}_i^i-\boldsymbol{M}_i\|_F^2\right)\\ &+\sum_{j=1}^{N}\left(\sum_{i=1}^{C}P_{i,j}\|\boldsymbol{b}_j-[\boldsymbol{D}_i\ \boldsymbol{E}_i]\boldsymbol{y}_j^i\|_F^2+\gamma\|\boldsymbol{y}_j^i\|_1\right)\\ &-\beta\left(-\sum_{j=1}^{N}\sum_{i=1}^{C}P_{i,j}\log P_{i,j}\right)\\ \text{s.t.}\quad &\boldsymbol{y}_j=\text{Code\_Classify}\left(\boldsymbol{b}_j,\boldsymbol{D},\boldsymbol{E}\right),\ \sum_{i=1}^{C}P_{i,j}=1\end{aligned} \tag{1}$$

where $\boldsymbol{X}_i^i$ and $\boldsymbol{y}_j^i$ are the coding coefficient matrix of $\boldsymbol{A}_i$ and unlabeled data $\boldsymbol{b}_j$ on the class-specific dictionary $[\boldsymbol{D}_i\ \boldsymbol{E}_i]$, respectively. The coding of unlabeled data, i.e., Code_Classify($\boldsymbol{b}_j$, $\boldsymbol{D}$, $\boldsymbol{E}$), is identical to the coding of testing data in the final classification. For simplicity, we make the regularization terms of both labeled coding vectors and unlabeled coding vectors share the same parameter of γ, which can make labeled data and unlabeled data share the same contribution for the model. All the parameters in Eq.(1) will be discussed in Section 5 " Experiments".

Analysis of DSSDL

a) discriminative representation

For the labeled training data, a discriminative representation term, *i.e.*, $\|\boldsymbol{A}_i-[\boldsymbol{D}_i\boldsymbol{E}_i]\boldsymbol{X}_i^i\|_F^2$ and a discriminative coefficient term, *i.e.*, $\|\boldsymbol{X}_i^i-\boldsymbol{M}_i\|_2^2$ are introduced, where $\boldsymbol{M}_i$ is the mean coefficient matrix with the same size as $\boldsymbol{X}_i^i$ and takes the mean column vector of $\boldsymbol{X}_i^i$ as its column vectors. Since both $\boldsymbol{D}_i$ and $\boldsymbol{E}_i$ are associated with i^{th} class, $\boldsymbol{A}_i$ should be well represented by $[\boldsymbol{D}_i\boldsymbol{E}_i]$ and the column vectors of $\boldsymbol{X}_i^i$ should be similar to each other. Because we want to enforce sparse representation for all class data, we should minimize the sum of $\|\boldsymbol{X}_i^i\|_1$.

For the unlabeled training data, a probability weighted data representation term for each class is required. For instance, a large $P_{i,j}$ (e.g., 1) indicates the j^{th} unlabeled train-

ing samples from i^{th} class, and the desired class-specific dictionary, *i.e.*, $[\boldsymbol{D}_i, \boldsymbol{E}_i]$, should well represented the j^{th} unlabeled training samples.

b) Identical coding of unlabeled data to final classification

In most of dictionary learning model, the coding strategy in training phase may not be consistent with that in the final classification, which cannot ensure enough discrimination of learned dictionary for the final classification. In our proposed DSSDL, we require that the coding strategy of unlabeled training data is the same to that of testing data in classification. Thus the discrimination learned in the training phase can be easily exploited in the final classification.

Different coding models of testing data are preferred, e.g., collaborative representation for face recognition and texture classification and local representation for digit recognition. To be identical to the coding in final classification, the collaborative representation coefficient of $\boldsymbol{b}_j$ is

$$\begin{aligned}&\text{Code\_Classify}\left(\boldsymbol{b}_j, \boldsymbol{D}, \boldsymbol{E}\right)\\&= \arg\min_{\boldsymbol{y}_j} \| \boldsymbol{b}_j - [\boldsymbol{D}\ \boldsymbol{E}]\boldsymbol{y}_j \|_F^2 + \gamma \| \boldsymbol{y}_j \|_1\end{aligned} \tag{2}$$

where $[\boldsymbol{D}\ \boldsymbol{E}] = [\boldsymbol{D}_1, \boldsymbol{E}_1, ..., \boldsymbol{D}_i, \boldsymbol{E}_i, ..., \boldsymbol{D}_C, \boldsymbol{E}_C]$ and $\boldsymbol{y}_j = [\boldsymbol{y}_j^1; ...\boldsymbol{y}_j^i; ...; \boldsymbol{y}_j^C]$.

The local representation of unlabeled training data is

$$\begin{aligned}&\text{Code\_Classify}\left(\boldsymbol{b}_j, \boldsymbol{D}, \boldsymbol{E}\right)\\&= \arg\min_{\boldsymbol{y}_j^i} \left\| \boldsymbol{b}_j - [\boldsymbol{D}_i\ \boldsymbol{E}_i]\boldsymbol{y}_j^i \right\|_F^2 + \gamma \| \boldsymbol{y}_j^i \|_1 \quad \forall\, i\end{aligned} \tag{3}$$

We can observe that in the coding phase of unlabeled training data, the coding coefficient is only related to the desired dictionary. That can avoid the impact of an incorrect estimation of $\boldsymbol{P}$.

C) Entropy regularization

The last term of Eq. (1) is an entropy regularization on the estimated class possibility of unlabeled data. Based on quite limited labeled training data, it is impossible to correctly estimate the identity of all unlabeled samples. Only minimizing $\sum_{i=1}^{C} P_{i,j} \| \boldsymbol{b}_j - [\boldsymbol{D}_i\ \boldsymbol{E}_i]\boldsymbol{y}_j^i \|_F^2$ would make $P_{i,j}$=1 if unlabeled sample $\boldsymbol{b}_j$ has the minimal reconstruction residual in i^{th}-class, which, however, may be a wrong estimation in practice. In order to reduce the risk of making the learned dictionary worse and worse, we introduce an entropy regularization on the estimated class probability of unlabeled training sample to better reflect the relationship between estimated class and unlabeled training data.

Optimization of DSSDL

The DSSDL objective function in Eq. (1) can be divided into two sub-problems by doing class estimation of unlabeled data and discriminative dictionary learning alternatively: updating $\boldsymbol{P}$ by fixing $\boldsymbol{D}$, $\boldsymbol{E}$ and $\boldsymbol{X}$, and updating $\boldsymbol{D}$, $\boldsymbol{E}$ and $\boldsymbol{X}$ by fixing $\boldsymbol{P}$.

Class estimation of unlabeled data

By fixing the class-particular dictionary and the coding coefficient (*i.e.*, $\boldsymbol{D}$, $\boldsymbol{E}$, $\boldsymbol{X}$ and $\boldsymbol{y}$), the DSSDL model of Eq. (1) becomes

$$\min_{\boldsymbol{P}} \begin{cases} \sum_{j=1}^{N}\sum_{i=1}^{C} P_{i,j} \| \boldsymbol{b}_j - [\boldsymbol{D}_i\ \boldsymbol{E}_i]\boldsymbol{y}_j^i \|_F^2 \\ -\beta\left(-\sum_{j=1}^{N}\sum_{i=1}^{C} P_{i,j} \log P_{i,j}\right) \end{cases} \quad \text{s.t.} \sum_{i=1}^{C} P_{i,j} = 1 \tag{4}$$

We update $\boldsymbol{P}$ sample by sample because the class estimation of unlabeled training data is independent to each other. Let $\varepsilon_j^i = \| \boldsymbol{b}_j - [\boldsymbol{D}_i \boldsymbol{E}_i]\boldsymbol{y}_j^i \|_F^2$. After some derivations illustrated in the supplementary material, and the class probability of the j^{th} unlabeled training data is

$$P_{i,j} = \exp\{-\varepsilon_j^i / \beta\} / \sum_{i=1}^{C} \exp\{-\varepsilon_j^i / \beta\} \tag{5}$$

Discriminative Dictionary learning

After the updating of probability matrix $\boldsymbol{P}$, more unlabeled training samples are chosen to train the discriminative dictionary. If we maintain the atom number of learnt dictionary, the discrimination of our dictionary cannot be fully utilized. Thus an extended dictionary atom $\boldsymbol{E}_i$ need to be initialized and added to sub-dictionary $\boldsymbol{D}_i$ in each iteration.

Initialization: Let $\hat{\boldsymbol{D}}_i$ denote the learnt i^{th}-class dictionary (i.e., include $\boldsymbol{D}_i$ and $\boldsymbol{E}_i$) in the last iteration. After some derivations in supplementary material, $\boldsymbol{E}_i$ is initialized as

$$\boldsymbol{E}_i = \boldsymbol{U}(:, n) \tag{6}$$

where n is the atom number of the extended dictionary. $[\boldsymbol{U}, \boldsymbol{S}, \boldsymbol{V}] = \text{svd}([(\boldsymbol{A}_i - \hat{\boldsymbol{D}}_i \boldsymbol{X}_i^i), \sqrt{P_{i,1}}\boldsymbol{\xi}_1^i, ..., \sqrt{P_{i,j}}\boldsymbol{\xi}_j^i, ..., \sqrt{P_{i,N}}\boldsymbol{\xi}_N^i])$ and $\boldsymbol{\xi}_j^i = \boldsymbol{b}_j - \hat{\boldsymbol{D}}_i \boldsymbol{y}_j^i$.

Dictionary updating: After obtaining the new extended dictionary (*i.e., $\boldsymbol{E}$)*, the discriminative dictionary learning (*i.e.*, updating $\boldsymbol{D}$, $\boldsymbol{E}$ and $\boldsymbol{X}$) in the DSSDL model of Eq. (1) becomes

$$\begin{aligned}&\min_{\boldsymbol{D},\boldsymbol{E},\boldsymbol{X}} \sum_{j=1}^{N}\left(\sum_{i=1}^{C} P_{i,j} \| \boldsymbol{b}_j - [\boldsymbol{D}_i\ \boldsymbol{E}_i]\boldsymbol{y}_j^i \|_F^2 + \gamma \| \boldsymbol{y}_j^i \|_1\right)\\&+\sum_{i=1}^{C}\left(\| \boldsymbol{A}_i - [\boldsymbol{D}_i\ \boldsymbol{E}_i]\boldsymbol{X}_i^i \|_F^2 + \gamma \| \boldsymbol{X}_i^i \|_1 + \lambda \| \boldsymbol{X}_i^i - \boldsymbol{M}_i \|_F^2\right)\end{aligned} \tag{7}$$

In the discriminative dictionary learning, the representation coefficient and dictionary are alternatively updated. When the dictionaries, $\boldsymbol{D}$ and $\boldsymbol{E}$, are known, the coding coefficient of labeled training data (e.g., for i-th class) can be easily updated via

$$\min_{\boldsymbol{X}} \| \boldsymbol{A}_i - [\boldsymbol{D}_i\ \boldsymbol{E}_i]\boldsymbol{X}_i^i \|_F^2 + \gamma \| \boldsymbol{X}_i^i \|_1 + \lambda \| \boldsymbol{X}_i^i - \boldsymbol{M}_i \|_F^2 \tag{8}$$

In our paper, we update $\boldsymbol{X}_i^i$ for i^{th}-class data by using the coding method in (Yang and Zhang 2014).

For the unlabeled training data, collaborative representation of Eq.(2) or local representation or Eq.(3) is conducted , depending on the specific classification task.

By fixing the coding coefficient, the dictionary can be updated class by class. the i^{th}-class dictionary is updated as

$$\min_{\boldsymbol{D}_i,\boldsymbol{E}_i} \| \boldsymbol{A}_i - [\boldsymbol{D}_i \ \boldsymbol{E}_i]\boldsymbol{X}_i^i \|_F^2 + \sum_{j=1}^{N} P_{i,j} \| \boldsymbol{b}_j - [\boldsymbol{D}_i \ \boldsymbol{E}_i]\boldsymbol{y}_j^i \|_F^2 \quad (9)$$

The dictionary updating of $[\boldsymbol{D}_i, \boldsymbol{E}_i]$ can be easily solved by using Metaface (Yang et al. 2010) with updating dictionary atom by atom.

Classification model of DSSDL

When the structured dictionary $[\boldsymbol{D}_i, \boldsymbol{E}_i]$ has been learnt for all classes, the classification procedure of DSSDL are described as follows.

1. Sparsely code the test image, $\boldsymbol{b}$, on the dictionary $[\boldsymbol{D}\ \boldsymbol{E}]=[\boldsymbol{D}_1, \boldsymbol{E}_1, ..., \boldsymbol{D}_i, \boldsymbol{E}_i, ..., \boldsymbol{D}_C, \boldsymbol{E}_C]$ via collaborative representation of Eq.(2) or local representation of Eq.(3). Let the coding coefficient be $\boldsymbol{y} = [\boldsymbol{y}^1; \boldsymbol{y}^2; \cdots; \boldsymbol{y}^C]$ and $\boldsymbol{y}^i$ is the coefficient vector associated with class i.

2. Classify $\boldsymbol{b}$ via

$$\text{identity}(\boldsymbol{b}) = \arg\min_i \{e_i\} \quad (10)$$

where $e_i = \|\boldsymbol{b} - [\boldsymbol{D}_i \ \boldsymbol{E}_i]\boldsymbol{y}^i\|_2^2$.

Experiments

We evaluate our approach on two face databases: Extended YaleB database (Lee, Jeffrey and David 2005), and LFW face database (Wolf, Hassner and Taigman 2009), two handwritten digit datasets: MNIST (LeCun et al. 1998) and USPS(Hull 1994) and an object category database: Texture(Lazebnik, Schimid and Ponce 2005) for the tasks of face recognition, digit recognition and texture classification, respectively. The competing methods include several representative supervised dictionary learning methods: SRC (Wright et al. 2009), M-SVM (Yang et al. 2009), FDDL(Yang and Zhang 2011), DKSVD(Zhang and Li 2010), LCKSVD(Jiang, Lin and Davis 2013), SVGDL(Cai et al. 2014), and semi-supervised dictionary learning methods: S2D2(Shrivastava 2012), JDL(Pham and Svetha 2008), OSSDL(Zhang, Jiang and Davis 2012), SSR-D(Wang et al. 2013), USSDL (Wang, Guo and Li 2015), and recently proposed SSP-DL(Wang et al. 2016) algorithm. As FDDL (Yang and Zhang 2011), the coding of unlabeled training data and testing data in the proposed DSSDL adopts local representation for digit recognition and collaborative representation for other experiments.

Parameter selection

There are three parameters, γ, λ, and β, in the model of DSSDL (i.e., Eq. (1)). γ is a parameter of sparse coding coefficient, λ is a parameter of discriminative coding coefficient term, and β is a parameter of maximum entropy. In our all experiments, we set γ=0.001 and λ=0.01 based on our experimental experience.

Parameter β is related to the weight of entropy. It cannot be too big since a strong entropy regularization makes the probabilities of unlabeled samples in different class similar, which leads to a poor classification performance. By the experimental experience, we set β=0.01 to suitably lower the weight of the unlabeled samples classified wrongly, while utilize the discrimination of learnt dictionary better.

Face recognition

The Extended YaleB database (Lee, Jeffrey and David 2005) consists of 2414 frontal face images of 38 individuals taken under varying illumination conditions. Each individual has 64 images and we randomly select 20 images as training set and use the rest as testing set. We follow the same experimental setting adopt in (Wang, Guo and Li 2015). And we randomly select {2, 5, 10} samples from each class in the training set as labeled data, and the remaining as the unlabeled data. The image samples are reduced to 300 dimension by PCA. The experiment were repeated 10 times to calculate the mean accuracy and standard deviation. Table 1 lists the face recognition results.

Table 1: The recognition rates (%)on Extended YaleB database.

Methods	2	5	10
SRC	47.8±2.9	79.9±1.9	90.0±0.5
M-SVM	38.0±2.6	66.6±1.1	83.8±0.8
FDDL	52.4±2.5	82.3±0.7	92.1±0.3
LC-KSVD	48.5±2.8	69.6±3.6	84.6±3.8
SVGDL	53.4±2.2	81.1±1.0	91.7±5.8
S2D2	53.4±2.1	76.1±1.3	83.2±1.9
JDL	55.2±1.8	77.4±2.8	85.3±1.6
USSDL	60.5±2.1	86.5±2.1	93.6±0.8
DSSDL	**62.1±3.0**	**87.5±0.3**	**94.5±0.3**

From the Table 1, it is clear that our proposed method achieves the highest recognition rates among the competing schemes. Especially when a small number of labeled samples are involved, the DSSDL performs crucially better than the supervised dictionary methods which are dependent on the number of the labeled samples. For example, when there are 2 labeled samples per class, DSSDL achieves about 10% higher recognition rate than FDDL, which is a state-of-the-art discriminative supervised dictionary learning method.

LFW (Wolf, Hassner and Taigman 2009) is a large-scale database, which contains variations of pose, illumination, expression, misalignment and occlusion. The 143 subjects with no less than 11 samples per subject are chosen (4174 images in total). For each person the first 10 samples are used for training data with the remaining samples for testing. We randomly select {3,5,7} samples from each class in the training set as labeled data, with the remaining traing data as unlabeled data. Histogram of Uniform-LBP is extracted via dividing a face image into 10×8 patches. Then we use PCA to reduce the histogram dimension to 500.

Table 2 illustrates the comparison of all methods (The performance of USSDL is not included in Table 2 since the code of USSDL is not publicly available). It can be seen that the mean recognition rate of DSSDL outperforms all the other semi-supervised methods by at least 2%. When there are only 3 labeled samples, 3% improvements are achieved by DSSDL compared to the supervised methods.

Table 2: The recognition rates (%) on LFW database.

Methods	3	5	7
SRC	42.7±0.6	55.2±1.9	62.2±2.7
DKSVD	39.6±0.8	50.3±0.7	56.7±1.8
LC-KSVD	41.8±1.0	49.2±0.9	58.6±1.3
FDDL	44.0±0.7	59.0±1.7	66.1±1.5
SVGDL	45.5±0.9	58.3±1.3	66.8±1.3
JDL	45.8±1.2	59.8±1.7	64.8±2.1
S2D2	46.2±1.8	61.0±1.8	65.4±2.1
DSSDL	**48.3±1.3**	**63.8±1.1**	**67.5±1.2**

Digit classification

We evaluate the performance on both the MNIST dataset and USPS dataset with the same experimental setting as (Wang et al. 2016). In the MNIST dataset, there are 10 classes and the training sets have 60,000 handwritten digital images and test sets have about 10,000 images. We randomly select 200 samples from each class. Then we select randomly 20 images as the labeled samples, 80 as the unlabeled samples and the rest used for testing. For USPS data sets, there are 9298 digital images consisting of 10 classes and we randomly select 110 images from each class. Then we randomly select 20 images as the labeled samples, 40 images as the unlabeled samples and 50 images as the testing samples. We use the whole image as the feature vector, and normalize the vector to have unit l_2-norm.

All relevant results for ten independent tests are listed in Table 3. It can be seen that the proposed DSSDL can effectively utilize information of the unlabeled samples, and the classification accuracy is at least 2.4% higher obviously than other dictionary methods. With the additional unlabeled training samples involved, the size of the dictionary is enlarged adaptively to better utilize the discrimination of the unlabeled samples. That is also why we can achieve better performance than other semi-supervised dictionary methods mentioned in the Table 3.

Table 3: The recognition rates (%) on USPS and Mnist.

Methods	USPS	Mnist
SRC	68.6±2.7	72.9±2.3
DKSVD	67.5±1.8	71.4±1.7
FDDL	85.2±1.2	82.5±1.3
LC-KSVD	76.9±1.3	73.0±1.3
OSSDL	80.8±2.8	73.2±1.8
S2D2	86.6±1.6	77.6±0.8
SSR-D	87.2±0.5	83.8±1.2
SSP-DL	87.8±1.1	85.8±1.2
DSSDL	**90.2±0.9**	**88.3±1.5**

Object classification

In this section, we take experiment on Texture-25 data set which contains 25 texture categories, 40 samples each. We use the low-level features (Boix, Roig, and Gool 2014; Dai and Gool 2013), including PHOG (Bosch, Zisserman and Munoz 2007), GIST (Oliva, Hospital and Ave 2001) and LBP (Ojala, Matti and Maenpaa 2002). Same to the experimental setting in (Wang, Guo and Li 2015), PHOG is computed with a 2-layer pyramid in 8 directions. GIST is computed on the rescaled images of 256 × 256 pixels, in 4, 8 and 8 orientations at 3 scales from coarse to fine. As for LBP, the uniform LBP is used. All the features are concatenated into a single 119-dimensional vector. In this experiment, 13 images are randomly selected for testing and randomly select {2, 5, 10, 15} samples from each class in the training set as labeled samples. The average accuracies together with the standard deviation in five independent tests are presented in Table 4.

It can be seen that DSSDL achieves at least 2% higher recognition rate than other all dictionary learning methods. When there are 15 labeled samples per class, the DSSDL achieves a better recognition rate close to 10% than USSDL, which gains the second best results. That is mainly because the powerful discrimination has been learned in the class-specific dictionary of DSSDL. JDL, which only uses the reconstruction error of both the labeled and unlabeled data, does not perform well.

Table 4. The recognition rates (%) on Texture25 database.

Methods	2	5	10	15
M-SVM	24.9±3.4	41.6±1.7	52.9±2.7	55.3±1.2
FDDL	31.4±4.0	48.9±1.7	52.6±3.1	56.7±1.4
LCKSVD	28.0±4.1	38.2±1.3	48.6±2.9	54.1±2.9
SVGDL	29.8±3.9	37.9±1.3	40.3±2.3	56.8±1.3
S2D2	31.7±2.3	43.8±1.4	47.9±2.4	50.9±1.7
JDL	27.6±2.1	39.2±1.9	43.3±0.8	50.3±0.8
USSDL	34.2±3.7	51.1±2.2	54.6±1.6	57.7±1.6
DSSDL	**36.8±3.4**	**55.3±3.6**	**64.3±3.7**	**68.2±1.6**

Conclusions

In this paper we proposed a discriminative semi-supervised dictionary learning model with an entropy regularization. The discrimination of labeled and unlabeled training data is explored by requiring discriminative representation residual and coefficients. For class estimation of unlabeled data, an entropy term is used to regularize their estimated probabilities. Meanwhile, an identical coding of unlabeled data to that of testing data is also required to ensure the learned discrimination suitable for classification. Extensive experiments on face recognition, digit recognition, and texture classification have shown its advantage over supervised dictionary learning methods and other semi-supervised dictionary learning approaches.

Acknowledgments

This work is partially supported by the National Natural Science Foundation for Young Scientists of China (no. 61402289), and National Science Foundation of Guangdong Province (no. 2014A030313558).

References

Aharon, M.; Michael, E.; Alfred, B. 2006. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE TSP*, 54(11):4311–4322.

Blum, A.; Mitchell. T. 1998. Combining labeled and unlabeled data with co-training. In *COLT*.

Bryt, O.; Elad, M. 2008. Compression of facial images using the k-svd algorithm. *IJCV*, 19(4): 270–282.

Cai, S.; Zuo, W.; Zhang, L.; Feng, X.; Wang, P. 2014. Support vector guided dictionary learning. In: *Proc. ECCV*.

Castrodad, A.; Sapiro, G. 2012. Sparse modeling of human actions from motion imagery. *IJCV*, 100(1):1–15.

Dai, D.; Gool, L.V. 2013. Ensemble projection for semi-supervised image classification. In: *Proc.ICCV*.

Elad, M.; Aharon, M. 2006. Image denoising via sparse and redundant representations over learnt dictionaries. In: *Proc. ICIP*.

Huang, K.; Aviyente, S. 2006. Sparse representation for signal classification. In: *Proc.NIPS*.

Hull, J. 1994. A database for handwritten text recognition research. *IEEE TPAMI*, 16(5): 550–554.

Jiang, Z.; Lin, Z.; Davis, L.S. 2013. Label consistent k-svd: Learning a discriminative dictionary for recognition. *IEEE TPAMI*, 35(11): 2651–2664, 2013.

Jian, M.; Jung, C. 2016. Semi-supervised bi-dictionary learning for image classification with smooth representation-based label propagation. *IEEE TM*, 18(3):458–473.

Kong, S.; Wang, D.H. 2012. A dictionary learning approach for classification: Separating the particularity and the commonality. In: *Proc. ECCV*.

Lee, K.-C.; Jeffrey, H.; David, K. 2005. Acquiring linear subspaces for face recognition under variable lighting. *IEEE TPAMI*, 27(5): 684–698.

LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

Lazebnik, S.; Schmid, C.; Ponce, J. 2005. A sparse texture representation using local affine regions. *IEEE TPAMI*, 27(8): 1265–1278.

Mairal, J.; Bach, F.; Ponce, J.; Sapiro, G.; Zisserman, A. 2008. Discriminative learnt dictionaries for local image analysis. In: *Proc. CVPR*.

Mairal, J.; Bach, F.; Ponce, J.; Sapiro, G.; Zisserman, A. 2009. Supervised dictionary learning. In: *Proc. NIPS*.

Mairal, J.; Bach, F.; Ponce, J. 2012. Task-driven dictionary learning. *IEEE TPAMI*, on 34(4): 791–804.

Mohamadabadi, B.; Zarghami, A.; Zolfghri, M.; Baghshah, S. 2013. Pssdl: Probabilistic semi-supervised dictionary learning. In: *Proc. ECML*.

Olshausen, B.A.; Field, D.J. 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381(6583): 607–609.

Pham, D.-S.; Svetha, V. 2008. Joint learning and dictionary construction for pattern recognition. In: *Proc. CVPR*.

Ramirez, I.; Sprechmann, P.; Sapiro, G. 2010. Classification and clustering via dictionary learning with structured incoherence and shared features. In: *Proc. CVPR*.

Rubinstein, R.; Bruckstein, A.M.; Elad, M. 2010. Dictionaries for sparse representation modeling. *Proceeding of the IEEE*, 98(6): 1045-1057.

Sindhwani, V.; Keerthi, S. S. 2006. Large scale semi-supervised linear svms. In: *ACM SIGIR*.

Shrivastava, A.; Pillai, J.K.; Patel, V.M.; Chellappa, R. 2012. Learning discriminative dictionaries with partially labeled data. In: *Proc. ICIP*.

Wang, X.; Guo, X.; Li, S. 2015. Adaptively unified semi-supervised dictionary learning with active points. In: *Proceeding of the ICCV*.

Wang, D.; Zhang, X.; Fan, M.; Ye, X. 2016. Semi-supervised dictionary learning via structural sparse preserving. In: *Proc. AAAI*.

Wang, H.; Nie, F.; Cai, W.; Huang, H. 2013. Semi-supervised robust dictionary learning via efficient l2,0+-norms minimization. In: *Proc. ICCV*.

Wright, J.; Yang, A.Y.; Ganesh, A.; Sastry, S.S.; Ma, Y. 2009. Robust face recognition via sparse representation. *IEEE TPAMI* 31(2): 210–227.

Wagner, A.; Wright, J.; Ganesh, A.; Zhou, Z.H.; Ma, Y. 2010. Towards a practical face recognition system: Robust registration and illumination by sparse representation. In: *Proc. ICASSP*.

Wolf, L.; Hassner, T.; Taigman, Y. 2009. Similarity Scores based on Background Samples. In: *Proc. ACCV*.

Yang, M.; Zhang, L.; Yang. J.; and Zhang, D. 2010. Metaface learning for sparse representation based face recognition. In: *Proc. ICIP*.

Yang, M.; Dai, D.; Shen, L.; Gool, L.V. 2014. Latent dictionary learning for sparse representation based classification. In: *Proc.CVPR*.

Yang, J.C.; Yu, K.; Gong, Y.; Huang, T.; 2009. Linear spatial pyramid matching using sparse coding for image classification. In: *Proc.CVPR*.

Yang, M.; Zhang, L.; Feng, X. 2011. Fisher Discrimination Dictionary Learning for Sparse Representation. In: *Proc.ICCV*.

Zhou, M.Y.; Chen, H.J.; Paisley, J.; Ren, L.; Li, L.B.; Xing, Z.M.; Dunson, D.; Sapiro, G.; Carin, L. 2010. Nonparametric bayesian dictionary learning for analysis of noisy and incomplete images. *IEEE TIP*, 21(1): 130– 144.

Zhang, Q.; Li, B.; 2010. Discriminative k-svd for dictionary learning in face recognition. In: *Proc. CVPR*.

Zhou, N.; Fan, J. 2012. Learning inter-related visual dictionary for object recognition. In: *Pro. CVPR*.

Zhang, G.; Jiang, Z.; Davis, L.S. 2012. Online semi-supervised discriminative dictionary learning for sparse representation. In: *Proc. ACCV*.

Zhu, X. 2005. Semi-supervised learning with graphs. In Proc. *IJCNLP*.