Computational Visual Media
DOI 10.1007/s41095-xxx-xxxx-x Vol. x, No. x, month year, xx–xx

Article Type (Research/Review)

Semi-supervised dictionary learning with label propagation for image classification

Lin Chen, Meng Yang ✉

Abstract Sparse coding and supervised dictionary learning has rapidly developed in recent years and achieved impressive performance in image classification. However, there are usually a limited number of labeled training samples and a huge number of unlabeled data in practical image classification, which degrades the discrimination of the learned dictionary. How to effectively utilize unlabeled training data and explore the discrimination hidden in unlabeled data has drawn much attention of researchers. In this paper, we propose a novel discriminative semi-supervised dictionary learning method by using label propagation (SSD-LP). Specifically, we utilize label propagation algorithm based on class-specific reconstruction errors to accurately estimate the identities of unlabeled training samples and develop an algorithm for optimizing the discriminative dictionary and discriminative coding vectors simultaneously. Extensive experiments on face recognition, digit recognition and texture classification demonstrate the effectiveness of the proposed method.

Keywords semi-supervised learning, dictionary learning, label propagation, image classification.

1 Introduction

In recent years, sparse representation has gained much interest in the computer vision field [13, 34] and been widely applied to image restoration [9, 35], image compression [6, 7] and image classification [23, 28, 39, 42, 44]. The success of sparse representation is partially because natural images can be generally and sparsely coded by structural primitives (e.g., edges and line segments) and the images or signals can be represented sparsely by the dictionary atoms from the same class.

In the task of image classification based on sparse representation, signals need to be encoded over a dictionary (i.e., a set of representation bases) with some sparsity constraint. The dictionary, which encodes the testing sample, can directly consist of the training samples themselves. For example, Wright et al.[36] firstly constructed a dictionary by using the training samples of all classes, then coded the test sample with this dictionary, and finally classified the test sample to the class with the minimal class-specific representation residual. The so-called sparse representation based classification (SRC) [36] has achieved impressive performance in face recognition. However, the number of the dictionary atoms used in SRC can be quite big, resulting in a high computational burden in calculating the coding vector. What is more, the discriminative information hidden in training samples cannot be exploited fully. To overcome the above problems, dictionary learning, i.e., how to learn an effective dictionary from training data, has been widely studied.

The dictionary learning methods can be divided into three main categories: unsupervised dictionary learning [1], supervised learning[17, 22, 38, 40] and semi-supervised dictionary learning[2, 16, 28, 29, 31, 33, 43]. K-SVD[1] is a representative unsupervised dictionary learning model, which is widely applied to image restoration tasks. Since no label information is exploited in the phase of dictionary learning, unsupervised dictionary learning methods are powerful for data reconstruction, but not advantageous for

1 College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. E-mail: chen.lin@email.szu.edu.cn.

2 College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. E-mail: yang.meng@szu.edu.cn.

Manuscript received: 2014-12-31; accepted: 2015-01-30.

classification tasks.

Based on the relationship between dictionary atoms and class labels, the prevailing supervised dictionary learning can be divided into three categories: shared dictionary learning, class-specific dictionary learning and hybrid dictionary learning. In the first case, discrimination of the shared dictionary learning was ordinarily explored by jointly learning a dictionary and a classifier over the coding coefficients [23, 44]. With the learned shared dictionary, the generated coding coefficients, which is expected to be discriminative, is used to conduct classification. In the class-specific dictionary learning, each dictionary atom is predefined to correspond to a unique class label so that the class-specific reconstruction error could be used for classification [22][41]. However, the size of the learned dictionary can be very big when there are a large number of classes. In order to utilize the powerful class-specific representation ability and reduce the coherence between different sub-dictionaries, the hybrid dictionary[12, 40, 45] combing shared dictionary atoms and class-specific dictionary atoms have been proposed.

Sufficient labeled training data and good quality of training images are necessary to improve the performance of supervised dictionary learning algorithm. However, it is expensive and difficult to obtain the labeled training data due to the vast human effort involved. On the other hand, there are abundant unlabeled images that can be collected easily from public image datasets. Therefore semi-supervised dictionary learning, i.e., how to effectively utilize these unlabeled samples to enhance dictionary learning, has attracted extensive researches.

In recent years, semi-supervised learning methods have been widely studied[3, 14, 24, 30, 46]. One classical semi-supervised learning method is co-training[44] which utilizes multi-view feature to retrain the classifiers to obtain better performance. In co-training, the multi-view feature need to be conditionally independent so that one classifier can select confident unlabeled samples for the other classifier. Another vital semi-supervised learning method is graph-based method[46]. In classification, graph-based semi-supervised learning methods can well explore the class information of unlabeled training data through a small number of the labeled data. A representative method based on graph is Label propagation (LP), which has been widely used in image classification and ranking. Label propagation algorithm[10, 18, 30, 37, 46] performs class estimation of unlabeled samples (i.e., label information propagations from labeled data to unlabeled data) by constructing a weight matrix (or affinity matrix) based on the distances between any two samples. The basic assumption of LP algorithm is that if the weight value between two samples is big, they are likely to belong to the same class.

Semi-supervised dictionary learning [2, 28, 29, 31–33, 43] has gained considerable interests in the past several years. In semi-supervised dictionary learning, whether the unlabeled samples can be accurately estimated and be used as labeled samples for training is very important. For instance, a shared dictionary and a classifier is jointly learned by estimating the class confidence of unlabeled samples [33]. In [2], the unlabeled samples are utilized to learn a discriminative dictionary by preserving the geometrical structure of both the labeled and unlabeled data. However, the class-specific reconstruction error which carries strong discriminative ability cannot be utilized to estimate the identities of unlabeled samples in shared dictionary. Semi-supervised class-specific dictionary has also been learned in [29]. However, its model is a little complex due to many regularizations.

By combining the label information of the labeled samples and reconstruction error of unlabeled samples over all classes, the identities of unlabeled training samples can be estimated more accurately. In this paper we propose a novel semi-supervised dictionary model with label propagation. In our proposed model, we design an improved label propagation algorithm to evaluate the probabilities of unlabeled data belonging to a specific class. Specifically, the proposed label propagation algorithm is based on the powerful class-specific representation (i.e., the reconstruction error of unlabeled samples on each sub-dictionary). Simultaneously, the label information of labeled data can be utilized better by this graph-based method, i.e., label propagation. We also well exploit the discrimination of labeled training data in dictionary learning by minimizing the within-class variance. We conducted several experiments on face recognition, digit recognition and texture classification, which shows the advantage of the proposed SSD-LP.

Our main contributions are summarized as follows:

1. We propose a novel discriminative semi-supervised dictionary learning method which can effectively utilize the discriminative information hidden in both the unlabeled and labeled training data.

2. By using the label propagation, we estimate a more accurate relationship between unlabeled training data and classes, and enhance the exploration of the

discrimination in unlabeled training data.

3. The discrimination of labeled training data by minimizing within-class variance is explored in the semi-supervised dictionary learning.

4. Experimental results show that our method has a significantly better discrimination ability with unlabeled training data used in the dictionary learning.

The rest of this paper is organized as follows. In Section 2, we briefly introduce the related work of semi-supervised dictionary learning. The proposed model is presented in Section 3. Section 4 describes the optimization procedure. Section 5 presents the experimental results and Section 6 concludes the paper with a brief summary and discussion.

2 Related work

Based on predefined relationship between dictionary atoms and class labels, semi-supervised dictionary learning can be mainly divided into two categories: discriminative class-specific dictionary learning and discriminative shared dictionary learning.

Motivated by [41], Shrivastava et al.[29] learnt a class-specific dictionary by using Fisher discriminant analysis on the coding vectors of the labeled data. However, its model is complex where the training data is represented on the combination of all class-specific dictionaries, and the coding coefficients are regularized by both intra-class and inter-class constraint.

Another type of semi-supervised dictionary learning is to learn a shared dictionary. Pham et al. [28] took into account the representation errors of both labeled data and unlabeled data. In addition, the classification errors of labeled data were incorporated into a joint objective function. One major drawback of the above approach is that it may encounter the problem of local minimum due to the dictionary construction and classifier design. Wang et al. [33] utilized an artificially designed penalty function to assign the weights to the unlabeled data and suppressed the weights of unlabeled data with low confidence much. Zhang et al. [43] proposed an online semi-supervised dictionary learning framework which integrated the reconstruction error of both labeled and unlabeled data, label consistency and the classification error into an objective function. Mohamadabadi et al. [2] integrated dictionary learning and classifier training into an objective function, and preserved the geometrical structure of both labeled and unlabeled data. Recently, Wang et al. [31] utilized the structural sparse relationships between both the labeled and unlabeled samples to learn a discriminative dictionary where the unlabeled samples can be automatically grouped to the different labeled samples. Although a shared dictionary usually has a compact size, the discrimination existing in class-specific reconstruction residuals cannot be used.

3 Semi-supervised dictionary learning with label propagation (SSD-LP)

Although several semi-supervised dictionary learning approaches have been proposed, there are still some issues needed to be solved, such as how to build a discriminative dictionary by using unlabeled data, how to utilize the representation ability of class-specific dictionary, and how to estimate the class possibilities of the unlabeled data. In this section, we propose a discriminative semi-supervised dictionary learning with label propagation (SSD-LP) method to address the issues mentioned above.

3.1 Model of SSD-LP

Let $\boldsymbol{A} = [\boldsymbol{A}_1, ...\boldsymbol{A}_i, ...\boldsymbol{A}_C]$ be the labeled 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 unlabeled training data with unknown labels from 1 and C, where N is the number of unlabeled training samples. Here similar to prevailing semi-supervised dictionary methods [2, 28, 29, 31–33, 43], our paper also assumes that the unlabeled training data belongs to some class of training set.

In our proposed model, the dictionary to be learnt is $\boldsymbol{D} = [\boldsymbol{D}_1, ..., \boldsymbol{D}_i, ..., \boldsymbol{D}_C]$, where $\boldsymbol{D}_i$ is the class-specific sub-dictionary associated with class i and is required to well represent i^{th}-class data but with a bad representation ability for all the other classes. In general, we make each column of $\boldsymbol{D}_i$ be a unit vector. We can write $\boldsymbol{D}_i$, the representation coefficient matrix of $\boldsymbol{A}_i$ over $\boldsymbol{D}$, as $\boldsymbol{X}_i = [\boldsymbol{X}_i^1; ...; \boldsymbol{X}_i^j; ...\boldsymbol{X}_i^C]$, where $\boldsymbol{X}_i^j$ is the coding coefficient matrix of $\boldsymbol{A}_i$ on the sub-dictionary $\boldsymbol{D}_j$. And $\boldsymbol{y}_j^i$ is the coding coefficient vector of the unlabeled sample $\boldsymbol{b}_j$ on the class-specific dictionary $\boldsymbol{D}_i$.

Apart from requiring the coding coefficients should be sparse, for the labeled training data we also minimize the within-class scatter of coding coefficients, $||\boldsymbol{X}_i^i - \boldsymbol{M}_i||$, to make the training samples from the same class have similar coding coefficients, 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.

We define a latent variable, $P_{i,j}$, which represents the probability of the j^{th} unlabeled training sample belonging to the i^{th} class. $P_{i,j}$ satisfies $0 \leq P_{i,j} \leq 1$ and $\sum_{i=1}^{C} P_{i,j} = 1$. For the labeled training sample k, if it belongs to class j, then $P_{j,k}=1$ and $P_{i,k}=0$ for $i\neq j$. Then the proposed SSD-LP can be formulated as

$$\min_{\boldsymbol{D},\boldsymbol{X},\boldsymbol{P},\boldsymbol{y}}\sum_{i=1}^{C}(||\boldsymbol{A}_i-\boldsymbol{D}_i\boldsymbol{X}_i^i||_F^2+\gamma||\boldsymbol{X}_i^i||_1-\lambda||\boldsymbol{X}_i^i-\boldsymbol{M}_i||_F^2)+\sum_{j=1}^{N}\left\{\sum_{i=1}^{C}P_{i,j}||\boldsymbol{b}_j-\boldsymbol{D}_i\boldsymbol{y}_j^i||_F^2+\gamma||\boldsymbol{y}_j^i||_1\right\}$$
$$s.t.\boldsymbol{P}\text{ is the output of Improved Label Propagation} \tag{1}$$

where γ and λ are parameter, and $\boldsymbol{P}$ is learned via the proposed Improved Label Propagation (ILP) algorithm.

For the labeled training data, a discriminative representation term, i.e., $||\boldsymbol{A}_i-\boldsymbol{D}_i\boldsymbol{X}_i^i||_F^2$ and a discriminative coefficient term, i.e., $||\boldsymbol{X}_i^i-\boldsymbol{M}_i||_2^2$ are introduced. Since $\boldsymbol{D}_i$ is associated with the i^{th}-class, it is expected that $\boldsymbol{A}_i$ should be well represented by $\boldsymbol{D}_i$ but not by $\boldsymbol{D}_j$, $j\neq i$. This implies that $\boldsymbol{X}_i^i$ should have some significant coefficients such that $||\boldsymbol{A}_i-\boldsymbol{D}_i\boldsymbol{X}_i^i||_F^2$ is small, while $\boldsymbol{X}_i^j$ should have nearly zero coefficients. Thus the term $||\boldsymbol{D}_i\boldsymbol{X}_i^j||_F^2$ is eliminated as shown in Eq.(1).

For the unlabeled training data, the probability that the sample belongs to each class is required. For instance, $P_{i,j}=1$ indicates the j^{th} unlabeled training sample from the i^{th}-class, and the class-specific dictionary $\boldsymbol{D}_i$ should well represent the j^{th} unlabeled training sample such that $||\boldsymbol{b}_j-\boldsymbol{D}_i\boldsymbol{y}_j^i||_F^2$ is small.

Due to the nice performance of graph-based label propagation in semi-supervised classification tasks, we utilize it to select the unlabeled sample with high confidence and assign the unlabeled sample a high weight, which is illustrated in detailed in subsection 4.1.

3.2 Classification scheme

Once the dictionary $\boldsymbol{D}$=$[\boldsymbol{D}_i,...,\boldsymbol{D}_i,...,\boldsymbol{D}_C]$ is learned, a testing sample can be classified via coding it over the learned dictionary. Although the learned dictionary is class-specific, the testing sample is not always coded on each sub-dictionary corresponding to each class. As discussion in [41], there are two methods of coding the testing sample.

When the number of training samples of each class is relatively small, the sample sub-space of class i cannot be supported by the learned sub-dictionary $\boldsymbol{D}_i$. Thus the testing samples $\boldsymbol{b}^t$ is represented on the collaborative combination of all class-specific dictionaries. In this case, the sparse coding vector of the testing sample should be gotten by solving

$$\hat{\boldsymbol{y}}=\arg\min_{\boldsymbol{y}}\{||\boldsymbol{b}^t-\boldsymbol{D}\boldsymbol{y}||_2^2+\gamma||\boldsymbol{y}||_1\} \tag{2}$$

where γ is a constant of sparsity constraint. Then the class of testing sample $\boldsymbol{b}^t$ is predicted by:

$$\text{label}=\arg\min_i||\boldsymbol{b}^t-\boldsymbol{D}_i\boldsymbol{y}_i||_2^2 \tag{3}$$

where $\hat{\boldsymbol{y}}=[\boldsymbol{y}_1;\boldsymbol{y}_2;...\boldsymbol{y}_i;...;\boldsymbol{y}_C]$ and $\boldsymbol{y}_i$ is the coefficient vector associated with class i.

When the number of training samples of each class is relatively large, the sub-dictionary $\boldsymbol{D}_i$, which has enough discrimination, can support the sample sub-space of class i. So we can directly code testing sample $\boldsymbol{b}^t$ on each sub-dictionary:

$$\hat{\boldsymbol{y}}=\arg\min_{\boldsymbol{y}}\{||\boldsymbol{b}^t-\boldsymbol{D}_i\boldsymbol{y}||_2^2+\gamma||\boldsymbol{y}||_1\} \tag{4}$$

Then the class of testing sample $\boldsymbol{b}^t$ is predicted by:

$$\text{label}=\arg\min_i\{e_i\} \tag{5}$$

where $e_i=||\boldsymbol{b}^t-\boldsymbol{D}_i\boldsymbol{y}_i||_2^2$, $\hat{\boldsymbol{y}}=[\boldsymbol{y}_1;\boldsymbol{y}_2;...\boldsymbol{y}_i;...;\boldsymbol{y}_C]$ and $\boldsymbol{y}_i$ is the coefficient vector associated with class i.

4 Optimization of SSD-LP

The SSD-LP objective function is not convex to the joint variable of $\{\boldsymbol{D},\boldsymbol{X},\boldsymbol{P},\boldsymbol{y}\}$, but it is convex to each variable when the others are fixed. The optimization of Eq.(1) is divided into three sub-problems: updating $\boldsymbol{P}$ by fixing $\boldsymbol{D}$, $\boldsymbol{X}$, $\boldsymbol{y}$; updating $\boldsymbol{X}$, $\boldsymbol{y}$ by fixing $\boldsymbol{P}$, $\boldsymbol{D}$; and updating $\boldsymbol{D}$ by fixing $\boldsymbol{P}$, $\boldsymbol{X}$.

4.1 Update P via Improved Label Propagation

Different from the way in [30] to construct the weight matrix, our weight matrix is constructed by the reconstruction errors of the unlabeled samples over all classes rather than the distances between any two samples. Intuitively, since sub-dictionary $\boldsymbol{D}_i$ have a good represent for i^{th}-class but have a bad represent for other classes, any pair of samples are likely to belong to the same class if they achieve the minimum reconstruction error in the same class.

Specifically, when we want to compute the weight value w_{ij}(if the weight value w_{ij} is big, then sample $\boldsymbol{b}_i$ and sample $\boldsymbol{b}_j$ are likely to have the same class), we firstly compute the reconstruction errors of both training sample $\boldsymbol{b}_i$ and $\boldsymbol{b}_j$ over all classes. Then we will get $\boldsymbol{e}_i$=$[e_{i1};e_{i2};...e_{ik};...e_{ic}]$ and $\boldsymbol{e}_j$=$[e_{j1};e_{j2};...e_{jk};...e_{jc}]$ where $e_{ik}=||\boldsymbol{b}_i-\boldsymbol{D}_k\boldsymbol{y}_i^k||_2^2$ denotes the reconstruction error value of sample $\boldsymbol{b}_i$ on the class k and $\boldsymbol{y}_i^k$ is the coefficient vector associated with class k.

After obtaining the $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$, we compute the distance d_{ij}^2 between $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$:

$$d_{ij}^2 = ||\boldsymbol{e}_i - \boldsymbol{e}_j||_2^2 \tag{6}$$

Finally, the weight value between sample $\boldsymbol{b}_i$ and sample $\boldsymbol{b}_j$ is:

$$w_{ij} = \exp(-\frac{d_{ij}^2}{\sigma^2}) \tag{7}$$

where σ is a constant. After counting all weight value between every two samples, we can get the transition matrix $\boldsymbol{T}$, which can be defined by normalizing the weight matrix as:

$$T(i,j) = \frac{w_{ij}}{\sum_{k=1}^{C} w_{ik}} \tag{8}$$

so that $\sum_{j=1}^{C} T(i,j) = 1$ and $\boldsymbol{T}$ is asymmetric after the normalization.

Define $n = n_l + n_u$ where n_l, n_u are the total number of the labeled training samples and unlabeled training samples, respectively. For the multi-class problem, the probability matrix is $\boldsymbol{P} = [\boldsymbol{P}^l; \boldsymbol{P}^u] \in \Re^{n*C}$, where C is the number of classes, $\boldsymbol{P}^l$ is the probability matrix for labeled samples, and $\boldsymbol{P}^u$ is the probability matrix for unlabeled samples. We let $\boldsymbol{P}^l(i,k) = 1$ if sample $\boldsymbol{b}_i$ is the labeled sample with class k, and 0 otherwise. We initialize the probability matrix as $\boldsymbol{P}_0 = [\boldsymbol{P}_0^l; \boldsymbol{0}]$, i.e., the probability of the unlabeled training samples are set as zero. The algorithm of improved label propagation in updating $\boldsymbol{P}$ is presented in Algorithm 1. The convergence of the Algorithm 1 can refer to the [46]. $\boldsymbol{P}_{t+1}$ denotes the next iteration of $\boldsymbol{P}_t$ and symbol * denotes the product of two matrices in the Algorithm 1. Please note that the step 3.b is crucial because it can always preserve the label information of the labeled sample.

Compared with the weight matrix based on the distances of any two original samples, there are two main advantages in our method. On one hand, original construction method of the weight matrix is a kind of single-track feedback mechanism where the update of the probability matrix $\boldsymbol{P}$ can affect the dictionary update, but the update of the latter cannot affect the former because the distances of original samples don't change. On the other hand, the weight matrix based on the reconstruction errors over all classes will more realistically reflect the similarity between two samples, which is helpful to estimate the class labels of unlabeled data.

4.2 Update X and y

By fixing the estimated class probabilities of the unlabeled training data (i.e., $\boldsymbol{P}$), the discriminative dictionary (i.e., $\boldsymbol{D}$) and coding coefficient (i.e., $\boldsymbol{X}$ and $\boldsymbol{y}$) are alternately updated.

Algorithm 1 Improved Label Propagation based on reconstruction error

1: Constructing a transition matrix $\boldsymbol{T}$ by Eq. (8);
2: Initializing the probability matrix $\boldsymbol{P}_0 = [\ \boldsymbol{P}_0^l\ ;\ \boldsymbol{0}]$;
3: Repeating the following steps until $\boldsymbol{P}$ converges:
  3.a $\boldsymbol{P}_{t+1}$=$\boldsymbol{T}$*$\boldsymbol{P}_t$;
  3.b $\boldsymbol{P}_{t+1}^l$=$\boldsymbol{P}_0^l$;
4: Output the probability matrix $\boldsymbol{P}$.

When the dictionary $\boldsymbol{D}$ is fixed, the coding coefficient of labeled training data can be easily updated. Now the objective function in Eq.(1) is reduced to:

$$\min_{\boldsymbol{X}} \sum_{i=1}^{C} (||\boldsymbol{A}_i - \boldsymbol{D}_i \boldsymbol{X}_i^i||_F^2 + \gamma ||\boldsymbol{X}_i^i||_1 + \lambda ||\boldsymbol{X}_i^i - \boldsymbol{M}_i||_F^2) \tag{9}$$

In our approach, we update $\boldsymbol{X}_i^i$ for i^{th}-class data by using the coding method in [41].

As discussed in section 3.2, when the number of training samples of each class are relatively small, updating coding coefficients of unlabeled training data with collaborative representation can achieve better classification performance. Inversely, we choose local representation when the training samples of each class are enough. For the unlabeled training data, two coding strategies, i.e., collaborative representation and local representation, are conducted. In the collaborative representation, the coding coefficient is solved via

$$\min_{\boldsymbol{y}_j} ||\boldsymbol{b}_j - \boldsymbol{D}\boldsymbol{y}_j||_F^2 + \gamma ||\boldsymbol{y}_j||_1 \tag{10}$$

where $\boldsymbol{D} = [\boldsymbol{D}_1, ..., \boldsymbol{D}_i, ..., \boldsymbol{D}_C]$ and $\boldsymbol{y}_j = [\boldsymbol{y}_j^1; ...; \boldsymbol{y}_j^i; ...; \boldsymbol{y}_j^C]$, $\boldsymbol{y}_j^i$ is the coding vector of the unlabeled sample $\boldsymbol{b}_j$ on the sub-dictionary $\boldsymbol{D}_i$. Here different class-specific dictionary,i.e., $\boldsymbol{D}_i$, will compete with each other in representing $\boldsymbol{b}_j$. In order to make a fair competition between different class-specific dictionaries, the encoding phase of collaborative representation ignores $\boldsymbol{P}$.

In the local representation, SSD-LP model associated to $\boldsymbol{y}_i$ changes to

$$\min_{\boldsymbol{y}_j^i} \sum_{i=1}^{C} P_{i,j}(||\boldsymbol{b}_j - \boldsymbol{D}_i \boldsymbol{y}_j^i||_F^2) + \gamma ||\boldsymbol{y}_j^i||_1 \tag{11}$$

which is a standard sparse coding problem.

4.3 Update D

After the updating of $\boldsymbol{P}$, more unlabeled training samples will be selected to train our model. If we maintain the atom number of learnt dictionary, the discrimination of our dictionary cannot be utilized better. Thus, after updating the probability matrix

$\boldsymbol{P}$, we should increase the size of each sub-dictionary to explore the discriminative information hidden in the unlabeled samples (i.e., an additional dictionary atom $\boldsymbol{E}_i$ need to be initialized and added to sub-dictionary $\boldsymbol{D}_i$).

Since the unlabeled samples introduce more discrimination, $\boldsymbol{E}_i$ is initialized by using unlabeled data in our paper.

$$\min_{\boldsymbol{E}} \sum_{j=1}^{N} \sum_{i=1}^{C} P_{i,j} ||\boldsymbol{b}_j - \boldsymbol{E}_i \tilde{y}_j^i||_F^2 \tag{12}$$

where $\tilde{y}_j^i$ is the unknown coding coefficient.

We update $\boldsymbol{E}_i$ class by class:

$$\min_{\boldsymbol{E}_i} \sum_{j=1}^{N} P_{i,j} ||\boldsymbol{b}_j - \boldsymbol{E}_i \tilde{y}_j^i||_F^2 \tag{13}$$

Then we combine all terms in Eq.(13):

$$\min_{\boldsymbol{E}_i} ||[\sqrt{P_{i,1}}\boldsymbol{b}_1, ..., \sqrt{P_{i,j}}\boldsymbol{b}_j, ..., \sqrt{P_{i,N}}\boldsymbol{b}_N] - \boldsymbol{E}_i[\sqrt{P_{i,1}}\tilde{\boldsymbol{y}}_1, ..., \sqrt{P_{i,j}}\tilde{\boldsymbol{y}}_j, ..., \sqrt{P_{i,N}}\tilde{\boldsymbol{y}}_N]||_F^2 \tag{14}$$

Since we require the coding coefficient should be sparse, we compute the extended dictionary by using singular-value decomposition (SVD):

$$[\boldsymbol{U}, \boldsymbol{S}, \boldsymbol{V}] = \mathrm{svd}((\sqrt{P_{i,1}}\boldsymbol{b}_1 ... (\sqrt{P_{i,j}}\boldsymbol{b}_j) ... \sqrt{P_{i,N}}\boldsymbol{b}_N)) \tag{15}$$

And the extended dictionary is defined that

$$\boldsymbol{E}_i = \boldsymbol{U}(:, n) \tag{16}$$

where n is the atom number of the extended dictionary. In all the experiment shown in this paper, we set n=1(i.e., each sub-dictionary will add an additional dictionary atom after the update of probability matrix $\boldsymbol{P}$).

Let the new sub-dictionary of class i be initialized as $\hat{\boldsymbol{D}}_i = [\boldsymbol{D}_i \ \boldsymbol{E}_i]$. By fixing the coding coefficient $\boldsymbol{X}$ and probability matrix $\boldsymbol{P}$, the objective function in Eq.(1) is reduced to:

$$\min_{\boldsymbol{D}} \sum_{i=1}^{C} (||\boldsymbol{A}_i - \hat{\boldsymbol{D}}_i \boldsymbol{X}_i^i||_F^2) + \sum_{j=1}^{N} (\sum_{i=1}^{C} P_{i,j} ||\boldsymbol{b}_j - \hat{\boldsymbol{D}}_i \boldsymbol{y}_j^i||_F^2) \tag{17}$$

The dictionary updating can be easily solved by using Metaface [42] with updating dictionary atom by atom. After updating the extended dictionary $\boldsymbol{E}$, we need several iterations to update the dictionary and coefficient, which can guarantee the convergence of the discriminative dictionary. In our experiment, the number of additional iteration is set to 5.

The whole algorithm of the proposed semi-supervised dictionary learning is summarized in Algorithm 2. The algorithm will converge since the total objective function value in Eq.(1) will decrease in each optimization. Figure 1 shows the total objective function value on AR dataset [25]. In all the experiments mentioned in this paper, our algorithm will converges in less than 10 iterations.

Algorithm 2 Semi-supervised dictionary learning with label propagation(SSD-LP)

1. Initialization
   The class probabilities of the unlabeled training samples are all initialized zero; the sub-dictionary $\boldsymbol{D}_i$ is initialized by the class i^{th}-class labeled training samples $\boldsymbol{A}_i$; and each column of $\boldsymbol{D}_i$ has unit l_2-norm.
2. Class estimation of the unlabeled training data
   Updating each $P_{i,j}$ for every unlabeled training sample in Algorithm 1.
3. Discriminative dictionary learning
   3.1 Computing the extended dictionary $\boldsymbol{E}_i$ for each class by solving Eq. (16).
   3.2 Updating the coding coefficient $\boldsymbol{X}$ and $\boldsymbol{y}$, and dictionary $\boldsymbol{D}$ for some iterations:
      3.2.1 Updating the coding coefficient by solving Eq. (9), (10) and (11);
      3.2.2 Updating the dictionary atom by atom by solving Eq. (17).
4. Return to step 2 until the values of the objective function in Eq. (1) in adjacent iterations are close enough or the maximum number of iterations is reached.
5. Output $\boldsymbol{D}$.

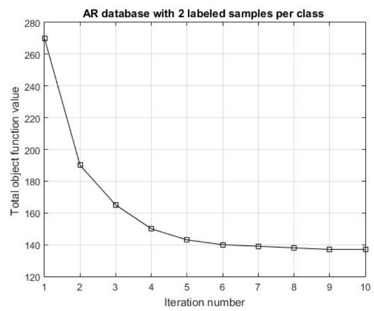

Fig. 1 The total objective function value on AR database [25].

5 Experiment results

We perform experiments and corresponding analysis to verify the performance of our method in image classification. We evaluate our approach on two face databases: Extended YaleB database [21] and the AR face database [25], two handwritten digit datasets: MNIST [20]and USPS [15], and an object category datasets: Texture-25 [19].We compare our method with SRC [36], M-SVM [38],FDDL [41], DKSVD [44], LCKSVD [17], SVGDL [8],S2D2 [29], JDL [28], OSSDL [43], SSR-D [32] and recently proposed USSDL [33] and SSP-DL [31] algorithm. The last six methods are semi-supervised dictionary learning models(*e.g.*, S2D2, JDL, OSSDL, SSR-D, USSDL and SSP-DL),and the others are supervised dictionary learning methods.

5.1 Parameter selection and compared with the original label propagation

In our all experiment, the parameters of SSD-LP chosen are fixed as γ=0.001 and λ=0.01. The additional iteration number is set to 5 in the step 3.b in the Tab. 2. In our experiment, since the sub-dictionary $\boldsymbol{D}_i$ is initialized by the i^{th}-class labeled samples, the atom number of $\boldsymbol{D}_i$ is equal to the number of the i^{th}-class labeled samples (e.g., in AR database, the number of sub-dictionary $\boldsymbol{D}_i$ are 2, 3, 5 when the labeled samples for each class are 2, 3, 5, respectively). After each update of probability matrix $\boldsymbol{P}$, each sub-dictionary will add an additional dictionary atom (the atom number of each sub-dictionary will not increase if the iteration exceed the number of unlabeled training number).

In order to show the effectiveness of our algorithm, a test on Extended Yale B is conducted. As shown in Figure 2, we can see that in the application of face recognition, the recognition significantly increases as the iteration number.

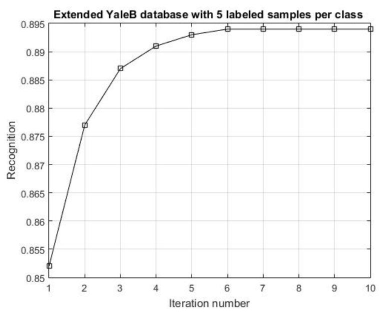

Fig. 2 The recognition rates versus iteration number on Extended YaleB database with five labeled training samples per class.

We also compare our proposed Improved Label Propagation with the original label propagation(LP). As shown in Figure 3, we can see that SSD-LP has the least 10% improvement than the performance where the images are classified directly by original label propagation. And with the increase of the number of iterations, the recognition rate of our method is growing but the performance of the original label propagation algorithm is unchanged basically. This is because the original label propagation is dependent on the distribution structure of input data which have no change with the update of dictionary so this is a kind of single-track feedback mechanism between the original label propagation and dictionary learning as discussion in the section 4.1.

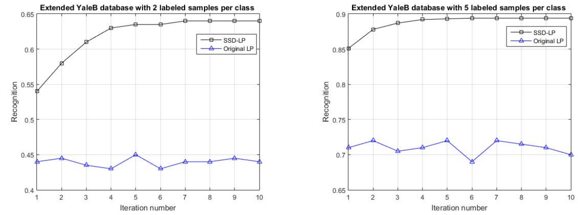

Fig. 3 The recognition rates versus iteration number on Extended YaleB database with two labeled training samples per person(at left) and five training samples per person(at right).

5.2 Face recognition

In this section, we evaluate our method on the experiment of face recognition on AR and Extended Yale B database with the same experimental setting as [33]. In both the two face recognition experiments, the image samples are reduced to 300 dimension by PCA. We firstly conduct on AR data sets. The AR database consists of over 4000 images of 126 individuals. In the experiment we choose a subset consisting of 50 male subjects and 50 female subjects. Focusing on the illumination and expression condition, for each subject we choose 7 images from Session 1 for training, and 7 images from Session 2 for testing. We randomly select 2, 3, 5 samples from each class in the training set as the labeled samples, and the remaining as the unlabeled samples. Five independent evaluations are conducted for the experiment with different number of the labeled training samples.

From the Tab. 1, when the number of the labeled samples are small such as two or three, our algorithm performs better than all the other methods, especially the supervised dictionary learning models. This is because supervised dictionary methods cannot utilize the discriminative information hidden in the unlabeled training samples. The semi-supervised dictionary learning methods usually perform better than the supervised dictionary learnings. For instance, USSDL performs the second best. Form the Table 1, we can see that USSDL has very close results to SSD-LP, but we should note that USSDL has more information in the dictionary learning task such as classifier learning of the coding vectors. In addition, the optimization procedure of USSDL is more complex than the SSD-LP.

We also evaluate our approach on Extended YaleB datasets. The database consists of 2414 frontal face images of 38 individuals. Each individual has 64 images and we randomly select 20 images as training set and use the rest as testing set. We randomly select 2, 5, 10

Tab. 1 The recognition rates (%) of all competing methods with different number of the labeled training samples on AR database

Methods	2	3	5
RC	72.2±1.0	79.1±0.9	88.2±0.5
M-SVM	60.1±2.2	74.3±1.2	84.9±2.0
FDDL	83.6±1.8	89.7±2.1	93.6±0.9
LC-KSVD	67.4±4.2	89.2±3.6	91.5±2.1
SVGDL	82.1±1.8	90.3±2.0	93.8±1.5
S2D2	85.3±3.1	89.2±1.9	92.1±1.1
JDL	87.2±2.0	88.2±1.8	90.7±1.2
USSDL	89.1±2.3	91.3±1.4	**94.1±1.3**
SSD-LP	**90.9±0.9**	**91.6±0.6**	93.7±0.5

samples from each class in the training set as the labeled samples, and the remaining as the unlabeled samples. And the classification results are shown in Tab. 2.

It is clear that our proposed method gains better classification performance than other dictionary learning methods. Especially when a small number of label samples are involved, the SSD-LP performs crucially better than the supervised dictionary learning methods which are dependent on the number of the labeled samples. It also can be seen that SSD-LP improves at least 1.5% over the other semi-supervised dictionary learning methods. When the number of labeled samples is small, the improvement will be more obvious. That is mainly because our method has strong capability to utilize the unlabeled samples by accurately determining their labels and using them as labeled samples to train our discriminative dictionary.

Tab. 2 The recognition rates (%) of all competing methods with different number of the labeled training samples on Extended YaleB database

Methods	2	3	5
SRC	47.8±2.9	79.1±1.9	90.5±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
SSD-LP	**67.0±2.9**	**89.8±0.9**	**95.2±0.2**

5.3 Digit classification

In this part, we evaluate the performance on both the MNIST dataset and USPS dataset with the same experimental setting as [31]. 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. In our experiment, the dimension of each digital image is 784 in MNIST dataset. We randomly select 200 samples from each class. And 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 l2-norm.

All relevant results for ten independent tests are combined in Table 3. It can be seen that the proposed SSD-LP can effectively utilize information of the unlabeled samples, and the classification accuracy is 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 why we can achieve better performance than other semi-supervised dictionary methods mentioned in the Tab. 3.

Tab. 3 The recognition rates (%) of all competing methods on two digit databases: 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
SSD-LP	**90.3±1.3**	**87.8±1.6**

5.4 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 [4, 11], including PHOG [5], GIST [27] and LBP [26]. Same to the experimental setting in [33], 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 Tab. 4.

It can be seen that SSD-LP improves at least 3% over the supervised dictionary learning when the number of labeled sample is 2 or 5. With the increase in the number of labeled samples, the effect is enhanced much obviously, about ten percent. From the Table 4, our method also have better results than the other three semi-supervised dictionary methods. That is because as adding more samples to train, the estimation to the label of unlabeled training data will be more accurate. The result fully demonstrates the classification effectiveness of label propagation based on reconstruction error. In addition, adaptively adding dictionary atoms makes our learnt dictionary more discriminative. JDL, which only uses the reconstruction error of both the labeled and unlabeled data, does not work well.

Tab. 4 The recognition rates (%) of all competing methods with different number of the labeled training samples on Texture25 database

Method	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
LC-KSVD	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
SSD-LP	**38.2±1.3**	**54.2±2.0**	**64.1±1.3**	**73.7±2.5**

6 Conclusions

In this paper we proposed a discriminative semi-supervised dictionary learning model. By integrating the label propagation with class-specific reconstruction error of each unlabeled training sample, we can more accurately estimate the class of unlabeled samples to train our model. The discrimination of labeled training data is also well explored by using a discriminative representation term and minimizing within-class scatter of the coding coefficients. Several experiments, including face recognition, digit recognition, and texture classification have shown its advantage over supervised dictionary learning methods and other semi-supervised dictionary learning approaches. In the future, we will explore more classification questions, *e.g.* the training samples may not belong to any known class.

Acknowledgements

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

References

[1] M. Aharon, M. Elad, and A. Bruckstein. k -svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006.

[2] B. Babagholami-Mohamadabadi, A. Zarghami, M. Zolfaghari, and M. S. Baghshah. Pssdl: Probabilistic semi-supervised dictionary learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 192–207. Springer, 2013.

[3] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. ACM, 1998.

[4] X. Boix, G. Roig, and L. Van Gool. Comment on" ensemble projection for semi-supervised image classification". *arXiv preprint arXiv:1408.6963*, 2014.

[5] A. Bosch, A. Zisserman, and X. Munoz. Image classification using random forests and ferns. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

[6] O. Bryt and M. Elad. Compression of facial images using the k-svd algorithm. *Journal of Visual Communication and Image Representation*, 19(4):270–282, 2008.

[7] O. Bryt and M. Elad. Improving the k-svd facial image compression using a linear deblocking method. In *2008 IEEE 25th Convention of Electrical and Electronics Engineers in Israel*, pages 533–537, 2008.

[8] S. Cai, W. Zuo, L. Zhang, X. Feng, and P. Wang. Support vector guided dictionary learning. In *European Conference on Computer Vision*, pages 624–639. Springer, 2014.

[9] Y.-C. Chen, V. M. Patel, P. J. Phillips, and R. Chellappa. Dictionary-based face recognition from video. In *European conference on computer vision*, pages 766–779. Springer, 2012.

[10] H. Cheng, Z. Liu, and J. Yang. Sparsity induced similarity measure for label propagation. In *2009 IEEE 12th International Conference on Computer Vision*,

pages 317–324, 2009.

[11] D. Dai and L. Van Gool. Ensemble projection for semi-supervised image classification. In *2013 IEEE International Conference on Computer Vision*, pages 2072–2079. IEEE, 2013.

[12] W. Deng, J. Hu, and J. Guo. Extended src: Undersampled face recognition via intraclass variant dictionary. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1864–1870, 2012.

[13] M. Elad, M. A. T. Figueiredo, and Y. Ma. On the role of sparse and redundant representations in image processing. *Proceedings of the IEEE*, 98(6):972–982, 2010.

[14] C. Gong, D. Tao, S. J. Maybank, W. Liu, G. Kang, and J. Yang. Multi-modal curriculum learning for semi-supervised image classification. *IEEE Transactions on Image Processing*, 25(7):3249–3260, 2016.

[15] J. J. Hull. A database for handwritten text recognition research. *IEEE Transactions on pattern analysis and machine intelligence*, 16(5):550–554, 1994.

[16] M. Jian and C. Jung. Semi-supervised bi-dictionary learning for image classification with smooth representation-based label propagation. *IEEE Transactions on Multimedia*, 18(3):458–473, 2016.

[17] Z. Jiang, Z. Lin, and L. S. Davis. Learning a discriminative dictionary for sparse coding via label consistent k-svd. In *CVPR 2011*, pages 1697–1704, 2011.

[18] F. Kang, R. Jin, and R. Sukthankar. Correlated label propagation with application to multi-label learning. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1719–1726, 2006.

[19] S. Lazebnik, C. Schmid, and J. Ponce. A sparse texture representation using local affine regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1265–1278, 2005.

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

[21] K.-C. Lee, J. Ho, and D. J. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on pattern analysis and machine intelligence*, 27(5):684–698, 2005.

[22] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Discriminative learned dictionaries for local image analysis. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[23] J. Mairal, J. Ponce, G. Sapiro, A. Zisserman, and F. R. Bach. Supervised dictionary learning. In *Advances in neural information processing systems*, pages 1033–1040, 2009.

[24] P. K. Mallapragada, R. Jin, A. K. Jain, and Y. Liu. Semiboost: Boosting for semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 31(11):2000–2014, 2009.

[25] A. M. Martinez. The ar face database. *CVC Technical Report*, 24, 1998.

[26] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987, 2002.

[27] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.

[28] D. S. Pham and S. Venkatesh. Joint learning and dictionary construction for pattern recognition. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[29] A. Shrivastava, J. K. Pillai, V. M. Patel, and R. Chellappa. Learning discriminative dictionaries with partially labeled data. In *2012 19th IEEE International Conference on Image Processing*, pages 3113–3116, 2012.

[30] B. Wang, Z. Tu, and J. K. Tsotsos. Dynamic label propagation for semi-supervised multi-class multi-label classification. In *2013 IEEE International Conference on Computer Vision*, pages 425–432, 2013.

[31] D. Wang, X. Zhang, M. Fan, and X. Ye. Semi-supervised dictionary learning via structural sparse preserving. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[32] H. Wang, F. Nie, W. Cai, and H. Huang. Semi-supervised robust dictionary learning via efficient l-norms minimization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1145–1152, 2013.

[33] X. Wang, X. Guo, and S. Z. Li. Adaptively unified semi-supervised dictionary learning with active points. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1787–1795, 2015.

[34] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan. Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE*, 98(6):1031–1044, 2010.

[35] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan. Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE*, 98(6):1031–1044, 2010.

[36] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227, 2009.

[37] C. Xiong and T. K. Kim. Set-based label propagation of face images. In *2012 19th IEEE International Conference on Image Processing*, pages 1433–1436, 2012.

[38] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *2009 IEEE Conference on Computer*

Vision and Pattern Recognition, pages 1794–1801, 2009.

[39] J. Yang, K. Yu, and T. Huang. Supervised translation-invariant sparse coding. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3517–3524, 2010.

[40] M. Yang, D. Dai, L. Shen, and L. V. Gool. Latent dictionary learning for sparse representation based classification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 4138–4145, 2014.

[41] M. Yang, L. Zhang, X. Feng, and D. Zhang. Fisher discrimination dictionary learning for sparse representation. In *2011 International Conference on Computer Vision*, pages 543–550, 2011.

[42] M. Yang, L. Zhang, J. Yang, and D. Zhang. Metaface learning for sparse representation based face recognition. In *2010 IEEE International Conference on Image Processing*, pages 1601–1604, 2010.

[43] G. Zhang, Z. Jiang, and L. S. Davis. Online semi-supervised discriminative dictionary learning for sparse representation. In *Asian Conference on Computer Vision*, pages 259–273. Springer, 2012.

[44] Q. Zhang and B. Li. Discriminative k-svd for dictionary learning in face recognition. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2691–2698, 2010.

[45] N. Zhou, Y. Shen, J. Peng, and J. Fan. Learning inter-related visual dictionary for object recognition. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3490–3497. IEEE, 2012.

[46] X. Zhu, J. Lafferty, and R. Rosenfeld. *Semi-supervised learning with graphs.* Carnegie Mellon University, language technologies institute, school of computer science, 2005.

Lin Chen Lin Chen received the B.S. degrees in Computer Science and Technology from Shenzhen University, Shenzhen, China, in 2015. He is currently pursuing the M.sc degree with the School of Computer Science & Software Engineering, Shenzhen University.

Meng Yang Meng Yang is currently an associate professor at School of Computer Science & Software Engineering, Shenzhen University, Shenzhen, China. He received his Ph.D degree from The Hong Kong Polytechnic University in 2012. Before joining Shenzhen University, he has been working as Postdoctoral fellow in the Computer Vision Lab of ETH Zurich. His research interest includes sparse coding, dictionary learning, object recognition and machine learning. He has published 10 AAAI/CVPR/ICCV/ICML/ECCV papers and several IJCV, IEEE TNNLS and TIP journal papers.