Joint learning and dictionary construction for pattern recognition

Duc-Son Pham and Svetha Venkatesh
Dept. of Computing, Curtin University of Technology
GPO Box U1987, Perth, WA 6845, Australia
dspham@ieee.org, svetha@cs.curtin.edu.au

Abstract

We propose a joint representation and classification framework that achieves the dual goal of finding the most discriminative sparse overcomplete encoding and optimal classifier parameters. Formulating an optimization problem that combines the objective function of the classification with the representation error of both labeled and unlabeled data, constrained by sparsity, we propose an algorithm that alternates between solving for subsets of parameters, whilst preserving the sparsity. The method is then evaluated over two important classification problems in computer vision: object categorization of natural images using the Caltech 101 database and face recognition using the Extended Yale B face database. The results show that the proposed method is competitive against other recently proposed sparse overcomplete counterparts and considerably outperforms many recently proposed face recognition techniques when the number training samples is small.

1. Introduction

Understanding natural images is a goal of much research in both computer vision and image processing. This understanding facilitates efficient coding algorithms in image processing and extraction of invariant features and classification for object categorization. Of interest is an recent important concept in these areas: sparse overcomplete representations. The work by Olshausen and Field [22, 23] highlights an interesting result from neuroscience that sparse overcomplete representations appear to be the underlying mechanism of the V1 sector of the primary visual cortex. Since then there have been a number of published works in this direction [18, 26]. Essentially, this spare overcomplete mechanism enables the extraction of invariant features that are well localized and suitable for classification. From the coding point of view, the mechanism effectively encodes or measures the similarity of a given pattern with a set of predefined templates, subject to transformations such as rotation, scaling, etc. Recently, cortex-like architectures for object recognition have been proposed with encouraging results [21, 25, 27].

Simultaneously, researchers in signal and image processing have investigated sparse overcomplete representations for a number of applications [8] under the concept *Sparse-Land* [9]. Examples of successful applications that achieve state-of-the-art performance include image denoising [9], image inpainting [19], and image compression [4]. Starting from the principle of linear superposition, it was found that overcomplete basis functions give several advantages over the complete counterpart, such as flexibility, robustness to noise, and most importantly, the number of basis function to represent the underlying signal or image is very small (i.e. sparse). A recent interesting result [12] on the equivalence between sparse approximation and the support vector machine indicates that sparsity is closely related to the number of support vectors. We note that though there exists a set of overcomplete basis functions for a particular class of images or signals, a randomly constructed overcomplete basis functions might not be well matched to the structure of the class [18]. This has led to recent research into the dictionary construction [1] as well as the theoretical study of the dictionary from an information theory perspective [8, 29]. Whilst the focus in this area is only on the representation aspect, it motivates us to exploit these advances in solving the classification problem.

We propose a new approach for pattern recognition using sparse overcomplete representations. Our approach is markedly different from previous work in that we propose to jointly construct the overcomplete dictionary and find the optimal classifier parameters This coincides with the idea of *deep learning*, essentially a joint source and channel coding from the information theory perspective [17]. In particular, we formulate a constrained optimization problem that involves the classification's objective function and the representation error of *both the labeled and unlabeled data*. Intuitively, this achieves the dual goal: reducing the regularized empirical risk under the statistical learning framework for the selection of the classifier [30], whilst maintaining small, overcomplete representation error with bounded

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

sparsity constraints. We propose an algorithm that achieves a suboptimal solution of the formulated optimization problem, by alternating between solving for subsets of parameters while preserving the sparsity. By including the unlabeled data in the above formulation, we seek to exploit the intrinsic information found across object categories that is otherwise limited by the labeled data, especially with small training samples. Our proposed method is evaluated over two important classification problems in computer vision: object categorization using the Caltech 101 database and face recognition using the Extended Yale B face database. The results show that the proposed method is competitive against other recently proposed sparse overcomplete counterparts using either ℓ_1-norm regularization [24] or cortex-like mechanism [27] over natural images. It also considerably outperforms many recently proposed face recognition techniques, especially when the number of training samples is small.

The novel aspect of our proposed frame work is that the sparse encoding coefficients are obtained in a *supervised* manner so as to match classification. In other words, the two stages work jointly, leading to the most *discriminative* sparse overcomplete representation that is suitable for the set of classifiers being considered. We note importantly, that this general principle has been seen in previous work, such as using cortex-like mechanism to obtain sparse encoding [27] or in the use of direct ℓ_1-norm regularization [24]. These methods, like the K-SVD algorithm [1], construct an overcomplete dictionary in an *unsupervised* manner, hence sparse overcomplete encoding in the representation stage is separated from finding the suitable classifier in the classification stage. Whilst this clearly helps simplify the task in each stage, the sparse overcomplete features extracted might not always be optimal in terms of *discriminative power* relative to the set of classifiers being considered.

The layout of this paper is as follows. Section II details our proposed method. Section III presents experimental results on the two classification problems. Concluding remarks are given in Section IV.

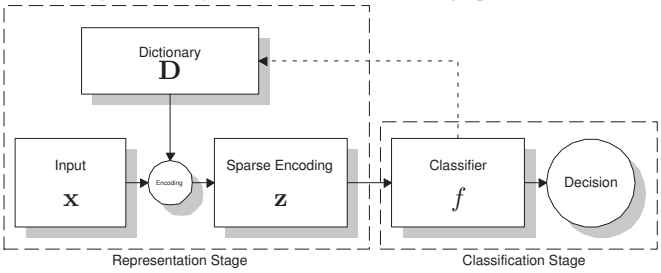
2. Proposed Method

Our starting point in this work is the availability of the following:

- A set of labeled images which we write in matrix form $\mathbf{X}_l = [\mathbf{x}_1^l, \ldots, \mathbf{x}_{N_l}^l]$ and unlabeled images $\mathbf{X}_u = [\mathbf{x}_1^u, \ldots, \mathbf{x}_{N_u}^u]$. Like many other approaches to create dictionary in an unsupervised manner, the set of unlabeled images enables the exploitation of intrinsic structure hidden within the class of natural images. We note that importantly in our formulation, the images can be the originals or pre-processed with low-level computer vision algorithms versions. Such pre-processing steps would certainly make the classification task easier.

- The embedding of the labels of the corresponding labelled images in a suitable space $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_{N_l}]$. This embedding is used to generalize a regression technique to the multivariate version [28] so that it can be directly used for classification. An example of the choice of embedding is to use with multivariate ridge regression [14] as a symmetric simplex in K-1 dimensions where K is the number of classes [2].

Figure 1. A system view of the two-stage process.



Our proposed method consists of two components that work jointly (see Fig. 1). First, in the representation stage, an input image $\mathbf{x}$ is converted to a sparse overcomplete representation $\mathbf{z}$. Next this sparse overcomplete representation is used for classification. In particular, we use a linear classifier for the second stage for its computational advantages. The objective functions in each stage are combined in one unified optimization problem so that a suitable sparse encoding strategy and the matching classifier can be jointly found.

The sparse overcomplete encoding is obtained under the availability of an overcomplete dictionary $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_k], \mathbf{d}_i \in \mathbb{R}^p$, which means that the number of columns is much more than the number of rows $k \gg p$. Each column of the dictionary, which is a unit norm vector is called an *atom*. The concept of dictionary and atoms are analogous to code book and code words in information theory. Each image $\mathbf{x}$ is a linear combination of the atoms in the dictionary, with the coefficients in vector $\mathbf{z}$. Ideally, we seek a representation such that the representation error $\| \mathbf{x} - \mathbf{D}\mathbf{z} \|^2$ is small, whilst ensuring the sparsity is bounded by $\| \mathbf{z} \|_0 < \epsilon$. However, it is known that such a direct solution will be computationally expensive. A more practical approach is to relax the original ℓ_0-norm regularization to a ℓ_1-norm regularization [8]

$$\mathbf{z} = \arg\min_{\mathbf{z}} \| \mathbf{x} - \mathbf{D}\mathbf{z} \|^2 + \lambda \| \mathbf{z} \|_1 . \quad (1)$$

Recent works have addressed the stability of the above relaxation and when such a relaxation can yield the exact ℓ_0-norm regularization. Essentially, this depends on the minimum distance between any two different atoms of $\mathbf{D}$, which

is known as the *coherence of the dictionary*, and the true sparsity. For details, please see [8, 29].

The source transform or sparse feature $\mathbf{z}$ is used for classification. Under the statistical learning framework we find the best classifier structure $f(\mathbf{z})$ from a set in the hypothesis space $\mathcal{H}$. Conventionally, the regularized version of the empirical risk is used to avoid overfitting and improve generalization ability, and thus

$$f = \arg\min_{f \in \mathcal{H}} \{R_{\text{emp}}[f] + R_{\text{reg}}[f]\}. \tag{2}$$

For simplicity, we consider the class of linear classifiers and in particular, the multivariate version $\mathbf{f}(\mathbf{z}; \mathbf{W}, \mathbf{b}) = \mathbf{W}^T\mathbf{z} + \mathbf{b}$, with the quadratic loss and ℓ_2-norm regularization. The above optimization is equivalent to the following problem:

$$[\mathbf{W}, \mathbf{b}] = \arg\min_{\mathbf{W},\mathbf{b}} \sum_{i=1}^{N_l} \parallel \mathbf{y}_i^l - \mathbf{W}^T\mathbf{z}_i^l - \mathbf{b} \parallel^2 + \gamma \parallel \mathbf{W} \parallel_F^2, \tag{3}$$

which is essentially a multivariate ridge regression problem. When the simplex of $\mathbf{y}_i$ is symmetric and the input is centralized, we can set the intercept $\mathbf{b} = \mathbf{0}$ and the ridge regression yields

$$\mathbf{W} = (\mathbf{Z}_l\mathbf{Z}_l^T + \gamma\mathbf{I})^{-1}\mathbf{Z}_l\mathbf{Y}_l^T \tag{4}$$

where $\mathbf{Z}_l = [\mathbf{z}_1^l, \ldots, \mathbf{z}_{N_l}^l]$ and $\mathbf{Y}_l = [\mathbf{y}_1^l, \ldots, \mathbf{y}_{N_l}^l]$ are the sparse encoding coefficients and multivariate labels of the training labeled data. We allow the dictionary $\mathbf{D}$ to be learnt to meet the dual goal:

- First, it should be adaptive to the pattern found in the data sets (for both labeled and unlabeled data). The K-SVD algorithm [1] is an example of finding such a suitable dictionary structure. When being adaptive to the data, the feature $\mathbf{z}$ is likely to be mostly sparse.
- Second, the dictionary $\mathbf{D}$ created should generate sparse overcomplete features $\mathbf{z}$ such that it carries the most discriminative information on the basis of the given hypothesis space $\mathcal{H}$ where the classifier structure is being specified.

It is noted that when one extracts discriminative features, the representation error may not necessarily be minimum. Such an example can be clearly seen with the difference between principal component analysis (PCA) and linear discriminant analysis (LDA).

When combined with the learning problem, our proposed approach leads to the following requirements

- The regularized empirical risk is small,
- The sparse overcomplete representation error is as small as possible but not necessary minimum, for a given upper bound on the sparsity of $\mathbf{z}$.

Though such a global solution of the optimization problem is difficult to obtain, we seek a local solution. To this end, we propose to combine the above requirements into a single optimization problem as:

$$\begin{aligned} \arg\min_{\mathbf{D},\mathbf{Z}} \quad & \parallel \mathbf{Y}_l - \mathbf{W}^T\mathbf{Z}_l \parallel_F^2 + \gamma \parallel \mathbf{W} \parallel_F^2 \\ & + \rho_l \parallel \mathbf{X}_l - \mathbf{D}\mathbf{Z}_l \parallel_F^2 \\ & + \rho_u \parallel \mathbf{X}_u - \mathbf{D}\mathbf{Z}_u \parallel_F^2 \\ \text{s.t.} \quad & \parallel \mathbf{z}_i \parallel_0 \leq \epsilon. \end{aligned} \tag{5}$$

The parameters ρ_l, ρ_u control the trade-off between the representation of the labeled and unlabeled data and classification. Loosely speaking, a large value of ρ places most emphasis on minimizing the representation error, whilst a small value of ρ would lose much representative ability. A suitable value of ρ would balance the goal of both objectives and improve classification performance. Another way to view of the above joint formulation is that the representation error terms act as the regularization on the parmeters $\mathbf{D}$ (hence $\mathbf{Z}$) of the classification problem. Though it is possible to set different values for the labeled and unlabeled data, in this work we simply consider the case $\rho_l = \rho_u = \rho$.

We also note that our formulation does not reduce the universality of a representative dictionary. Indeed, our method starts from a universal representative dictionary that can be used for a wide range of image classes. It then obtains a discriminative dictionary for a specific classification task by gradually altering the entries in a representative dictionary under sparsity constraints and to minimize the regularised risk functions. Our method is also general in a sense that when setting the values of ρ_l and ρ_u to ∞ we obtain ordinary methods directly using a representative dictionary. We also note that a previous work [15] also attempts to adjust the trade-off between sparsity and Fisher discrimination power by setting up a related optimization problem. However, their formulation uses a fixed representative dictionary, hence it restricts them from exploiting the useful information found in the unlabeled data. When there are few training samples, their approach is clearly limited as it is more likely to lose generalization ability.

Solving (5) is a challenging task and it appears that such a global solution might not be analytically available. What we propose in the following is an iterative algorithm that alternates between variables such that the updates are tractable.

Algorithm.

- **Step 1:** *Initialize the dictionary* $\mathbf{D}$ *and spare encoding coefficients* $\mathbf{Z}$ *using the K-SVD algorithm so that it satisfies the sparsity constraints.*
- **Step 2:** *Fix the dictionary* $\mathbf{D}$ *and* $\mathbf{Z}$ *and learn the linear classifier, i.e.* $\mathbf{W}$*, using (4).*

- **Step 3:** *Fix* $\mathbf{W}$ *and update the dictionary* $\mathbf{D}$ *and the sparse coefficients* $\mathbf{Z}$ *in a atom-by-atom fashion such that the regularized empirical risk is further reduced whilst making sure of good sparse representation (delineated below).*

- **Step 4:** *Check for convergence of the objective function in (5), otherwise repeat Steps 2 and 3.*

All steps in the above algorithm are obvious except for Step 3 which we explain next. In Step 3, as we fix $\mathbf{W}$ it reduces to solving:

$$\begin{aligned} \arg\min_{\mathbf{D},\mathbf{Z}} \quad & \| \mathbf{Y}_l - \mathbf{W}^T\mathbf{Z}_l \|_F^2 \\ & +\rho_l \| \mathbf{X}_l - \mathbf{D}\mathbf{Z}_l \|_F^2 \\ & +\rho_u \| \mathbf{X}_u - \mathbf{D}\mathbf{Z}_u \|_F^2 \\ \text{s.t.} \quad & \| \mathbf{z}_i \|_0 \leq \epsilon. \end{aligned} \tag{6}$$

Following the spirit of the K-SVD algorithm, we shall update each atom $\mathbf{d}_i$ of the dictionary $\mathbf{D}$ at a time and its corresponding sparse encoding, so that sparsity is under control. Suppose that we are updating the atom $\mathbf{d}_i$, we first rewrite:

$$\mathbf{D} = [\mathbf{D}_1 \ \mathbf{d}_i \ \mathbf{D}_2], \tag{7}$$

$$\mathbf{Z}^T = [\mathbf{R}_1 \ \mathbf{r}_i \ \mathbf{R}_2], \tag{8}$$

so that the second and third terms of (6) can be generally expressed in the following form:

$$\| \mathbf{X} - \mathbf{D}\mathbf{Z} \|_F^2 = \| \mathbf{E} - \mathbf{d}_i\mathbf{r}_i^T \|_F^2 \tag{9}$$

where $\mathbf{E} = \mathbf{X} - \mathbf{D}_1\mathbf{R}_1^T - \mathbf{D}_2\mathbf{R}_2^T$. Note that each row $\mathbf{r}_i^T$ involves many ith entries of the columns of $\mathbf{Z}$. Due to sparsity, many entries of $\mathbf{r}_i^T$ are also zero[1]. Let $\tilde{\mathbf{r}_i}^T$ denote the result of removing zero entries in $\mathbf{r}_i^T$ and the corresponding effect on $\mathbf{E}$ is $\tilde{\mathbf{E}}$. Then,

$$\| \mathbf{X} - \mathbf{D}\mathbf{Z} \|_F^2 = \| \tilde{\mathbf{E}} - \mathbf{d}_i\tilde{\mathbf{r}}_i^T \|_F^2 + \text{const} \tag{10}$$

where the constant is with respect to the optimization over $\mathbf{d}_i$. Note that Equation (10) equally applies to both the labeled and unlabeled data and we have removed the subscript for notational simplicity.

Next, we simplify the term for the classifier. In a similar fashion, let $\mathbf{W}^T = [\mathbf{C}_1 \ \mathbf{c}_i \ \mathbf{C}_2]$ and let $\mathbf{H} = \mathbf{Y}_l - \mathbf{C}_1\mathbf{R}_{l1}^T - \mathbf{C}_2\mathbf{R}_{l2}^T$ then

$$\| \mathbf{Y}_l - \mathbf{W}^T\mathbf{Z}_l \|_F^2 = \| \mathbf{H} - \mathbf{c}_i\mathbf{r}_{li}^T \|^2 . \tag{11}$$

After discarding the zero entries in $\mathbf{r}_i$, we have

$$\| \mathbf{Y}_l - \mathbf{W}^T\mathbf{Z}_l \|_F^2 = \| \tilde{\mathbf{H}} - \mathbf{c}_i\tilde{\mathbf{r}}_{li}^T \|_F^2 + \text{const}. \tag{12}$$

[1]We note importantly that in practice some ℓ_1 solver will generate truly sparse solution, i.e. many entries are zero, but some solvers generate approximately sparse solution, i.e. many entries are not zero but very small in magnitude. In that case, one can discard entries whose magnitude is less than a certain threshold.

Using (12) and (10), to update atom $\mathbf{d}_i$ and the corresponding sparse encoding in $\mathbf{Z}_l$ and $\mathbf{Z}_u$, we need to solve:

$$\begin{aligned} \min_{\mathbf{d}_i,\tilde{\mathbf{r}}_{li},\tilde{\mathbf{r}}_{ui}} \quad & \| \tilde{\mathbf{H}} - \mathbf{c}_i\tilde{\mathbf{r}}_{li}^T \|^2 + \rho_l \| \tilde{\mathbf{E}}_l - \mathbf{d}_i\tilde{\mathbf{r}}_{li}^T \|_F^2 \\ & +\rho_u \| \tilde{\mathbf{E}}_u - \mathbf{d}_i\tilde{\mathbf{r}}_{ui}^T \|_F^2 . \end{aligned} \tag{13}$$

Note that due to the nature of discarding the zero entries in $\mathbf{Z}$, its sparsity is always under control. We first consider the case when $\| \mathbf{c}_i \| = 0$. In this case, the first term becomes a constant with respect to the optimization problem. Hence it reduces to a standard rank-1 approximation

$$\begin{aligned} \min_{\mathbf{d}_i,\tilde{\mathbf{r}}_{li},\tilde{\mathbf{r}}_{ui}} \| \begin{bmatrix} \sqrt{\rho}_l\tilde{\mathbf{E}}_l & \sqrt{\rho}_u\tilde{\mathbf{E}}_u \end{bmatrix} - \\ \mathbf{d} \begin{bmatrix} \sqrt{\rho}_l\tilde{\mathbf{r}}_{li}^T & \sqrt{\rho}_u\tilde{\mathbf{r}}_{ui}^T \end{bmatrix} \|_F^2 . \end{aligned} \tag{14}$$

More precisely, suppose that the SVD of

$$\begin{aligned} \mathbf{\Xi} &= \begin{bmatrix} \sqrt{\rho}_l\tilde{\mathbf{E}}_l \\ \sqrt{\rho}_u\tilde{\mathbf{E}}_u \end{bmatrix} \\ &= \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \end{aligned} \tag{15}$$

then $\mathbf{d}$ is the eigenvector corresponding to the largest eigenvalues and $\begin{bmatrix} \sqrt{\rho}_l\tilde{\mathbf{r}}_{li}^T & \sqrt{\rho}_u\tilde{\mathbf{r}}_{ui}^T \end{bmatrix}$ is equal to the corresponding row of $\mathbf{V}^T$ multiplied by the largest eigenvalue.

To solve for the case when $\| \mathbf{c}_i \| \neq 0$, we start with the following result

Lemma 1 *The solution of the optimizations problems:*

$$\mathcal{P}_1 : \min_{\mathbf{b}\in\mathbb{R}^k} \| \mathbf{M} - \mathbf{a}\mathbf{b}^T \|_F^2 \tag{16}$$

$$\mathcal{P}_2 : \min_{\mathbf{a}\in\mathbb{R}^n,\mathbf{a}^T\mathbf{a}=1} \| \mathbf{M} - \mathbf{a}\mathbf{b}^T \|_F^2 \tag{17}$$

are given as follows:

$$\mathbf{b} = \frac{1}{\| \mathbf{a} \|^2}\mathbf{M}^T\mathbf{a}, \tag{18}$$

$$\mathbf{a} = \frac{1}{\| \mathbf{M}\mathbf{b} \|}\mathbf{M}\mathbf{b} \tag{19}$$

provided that either $\| \mathbf{a} \| \neq \mathbf{0}$ *or* $\| \mathbf{M}\mathbf{b} \| \neq \mathbf{0}$

A proof of this result is given in Appendix I. Now, we come back to problem (13) when $\| \mathbf{c}_i \| \neq 0$. We note that if we fix $\tilde{\mathbf{r}}_{li}$, and $\mathbf{d}_i$, then the problem is equivalent to

$$\min_{\tilde{\mathbf{r}}_{ui}} \| \tilde{\mathbf{E}}_u - \mathbf{d}_i\tilde{\mathbf{r}}_{ui}^T \|_F^2 . \tag{20}$$

On the other hand, if we fix $\tilde{\mathbf{r}}_{ui}$ and $\mathbf{d}_i$, then the problem is equivalent to

$$\min_{\tilde{\mathbf{r}}_{li}} \| \begin{bmatrix} \tilde{\mathbf{H}} \\ \sqrt{\rho_l}\tilde{\mathbf{E}}_l \end{bmatrix} - \begin{bmatrix} \mathbf{c}_i \\ \sqrt{\rho_l}\mathbf{d}_i \end{bmatrix} \tilde{\mathbf{r}}_{li}^T \|_F^2 . \tag{21}$$

Finally, if we fix $\tilde{\mathbf{r}}_{li}$ and $\tilde{\mathbf{r}}_{ui}$, the problem reduces to

$$\min_{\mathbf{d}_i} \| \begin{bmatrix} \sqrt{\rho_l}\tilde{\mathbf{E}}_l \\ \sqrt{\rho_u}\tilde{\mathbf{E}}_u \end{bmatrix} - \mathbf{d}_i \begin{bmatrix} \sqrt{\rho_l}\mathbf{r}_{li}^T \\ \sqrt{\rho_u}\mathbf{r}_{ui}^T \end{bmatrix} \|_F^2 . \quad (22)$$

Together with Lemma 1, these observations suggest that we can alternate between solving for $\tilde{\mathbf{r}}_{ui}$, $\tilde{\mathbf{r}}_{li}$, and $\mathbf{d}_i$, so that Step 3 of the proposed algorithm can be split into three iterated sub-steps:

- **Step 3a:** We solve for $\tilde{\mathbf{r}}_{ui}$ while fixing $\tilde{\mathbf{r}}_{li}$, and $\mathbf{d}_i$. To do so, we apply the result (18) for $\mathbf{M} = \tilde{\mathbf{E}}_u$, $\mathbf{b} = \tilde{\mathbf{r}}_{ui}$ and $\mathbf{a} = \mathbf{d}_i$.

- **Step 3b:** We solve for $\tilde{\mathbf{r}}_{li}$ while fixing $\tilde{\mathbf{r}}_{ui}$ and $\mathbf{d}_i$. To do so, we apply the result (18) for

$$\mathbf{M} = \begin{bmatrix} \sqrt{\rho_l}\tilde{\mathbf{E}}_l \\ \tilde{\mathbf{H}} \end{bmatrix} \quad (23)$$

$$\mathbf{a} = \begin{bmatrix} \sqrt{\rho_l}\mathbf{d} \\ \mathbf{c}_i \end{bmatrix} \quad (24)$$

  and $\mathbf{b} = \tilde{\mathbf{r}}_{li}$.

- **Step 3c:** We solve for $\mathbf{d}_i$ while fixing $\tilde{\mathbf{r}}_{li}$ and $\tilde{\mathbf{r}}_{ui}$. To do so, we apply the result (19) for

$$\mathbf{M} = \begin{bmatrix} \sqrt{\rho_l}\tilde{\mathbf{E}}_l \\ \sqrt{\rho_u}\tilde{\mathbf{E}}_u \end{bmatrix} \quad (25)$$

$$\mathbf{b} = \begin{bmatrix} \sqrt{\rho_l}\tilde{\mathbf{r}}_{li} \\ \sqrt{\rho_u}\tilde{\mathbf{r}}_{ui} \end{bmatrix} \quad (26)$$

  and $\mathbf{a} = \mathbf{d}_i$.

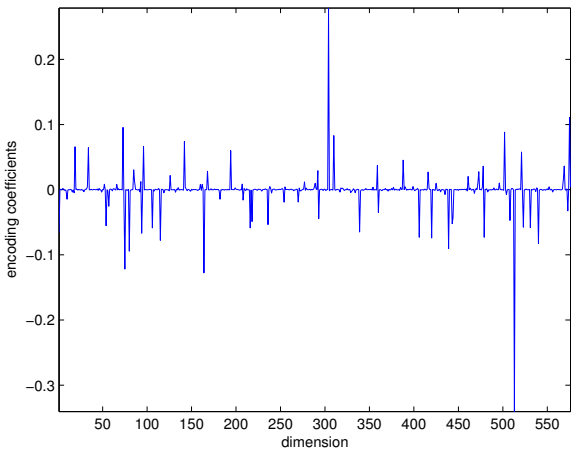
3. Experimental Results

In this section, we demonstrate the applicability of our framework in two major image classification problems, namely general object categorization and face recognition. Whilst general object categorization deals with a large variety of classes, each of which can have a large variety of shapes and textures, face recognition deals with a particular type of image (face), but has different individuals. The selection of these two problems is to demonstrate the wide applicability of our proposed supervised dictionary learning framework to various tasks.

3.1. Object categorization

The most widely used, yet challenging database for object recognition is the Caltech 101 [10]. This dataset has a total of 9,144 different images over 101 categories with an additional background class. The categories here are not totally separable in some sense. For example, there are categories referring to the whole animal body whilst there also exist other categories that have only the animal heads. Another example is the category called `faces_easy` and the other category `faces` which differ mostly in the fact one occupies most of the image region whilst the other only occupies a small region. The background category contains a wide range of items. The images have variable aspect ratio with an average size of about 300 pixels each dimension. The number of images in each category also varies from 31 (`inline_skate`) to 800 (`airplanes`). Following the standard testing procedure in many works on this dataset, we select randomly 30 images from each category and use up to 15 random images for training and the rest for testing. As our framework does not deal with the scaling issue, we assume such pre-processing steps are available. In practice, this can be justified using suitable segmentation techniques to set a bounding box containing the object of interest. In the experiment, we manually crop the region of interest and convert the images to gray scale, use histogram normalization, resize to 32×32, and use dimensionality reduction technique (PCA) to convert the images to unit vectors of size 144. We remove the background and faces categories, thus using a total of 100 categories.

Figure 2. Typical sparse overcomplete encoding coefficients.



The first step in the experiment is to obtain a initial estimate of the dictionary. This step is unsupervised and hence does not require the availability of class labels. We use the standard K-SVD algorithm to perform this task. To obtain the sparse overcomplete encoding coefficients $\mathbf{z}_i$ described in (1), we use the ℓ_1 solver from [16] and set the regularization parameter $\lambda = 0.05$[2]. The redundancy factor for the dictionary is set to 4 (i.e. the number of atoms in $\mathbf{D}$ is 4 times the dimension of each atom), which is hinted at from previous work [4]. In this initialization step, we take all im-

[2]The choice of this value is clearly not optimal and only based on our observation that this would lead to a reasonably sparse result.

ages in the database to compute the initial dictionary[3]. A typical sparse overcomplete encoding example is illustrated in Fig. 2. The rest of the experiment is conducted over a total of 20 random splits of the data. In each split, for simplicity, we take the testing data as unlabeled data[4] and set the parameter $\rho_l = \rho_u = 10^4$. We measure the average error per class over these 20 runs as well as the standard deviation.

On the Caltech 101 database, for 15 training samples, the average classification accuracy is $42\% \pm 1\%$. The confusion matrix is shown in Fig. 3. The classes receive highest accuracy are `yin_yang`,`faces_easy` and `dollar_bill` whilst lowest accuracy is observed over `ant`, `leopards`, and `llama`. The most confusing pair is `mayfly` and `saxophone`. Compared with recent work using simple ℓ_1-norm regularization approach [24] with the result for 1-region of just 30%, our result is encouraging. It is important to note that [24] and many other approaches use sophisticated classifier structures whilst we only use a simple linear classifier. The recently published work using the cortex type approach reports an average accuracy of 43% [27]. The highest recognition accuracy of 79.85% was reported in [31] which uses sophisticated low-level vision techniques and complex classifier. Our results are attractive as both the choice of the classifier and pre-processing is markedly simpler. The other point to mention is that other works also assume high resolution images, whilst we only work with a size of 32×32 and then convert them to a vector of 144, which is much faster and more suitable for applications such as in video surveillance. We envisage that it is possible to extend our method to larger images in a similar spirit to [20] so that performance improvement can be achieved.

3.2. Face recognition

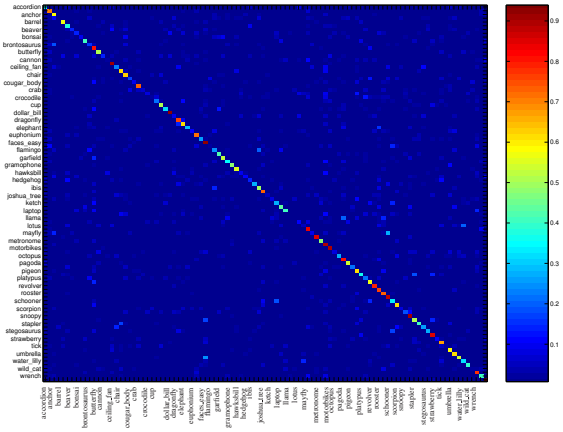
We select a set of 2414 near frontal images for 38 individuals from Extended Yale B face database [11] for this experiment[5]. This database addresses the issue of pose and illumination, which are difficult problems in face recognition. Our pre-processing step is similar to the previous experiment: the original images are cropped to 32×32 pixels, converted to column vectors, normalized to unit norm, then projected to a dimension of 144 using PCA. We divide the database into two disjoint subsets, each containing 19 individuals. One is used for the *labeled* data and testing, whilst

[3]Even with the K-SVD itself, one needs to specify the initial value of the dictionary, for example using an overcomplete DCT version. In this experiment, we decide to initialize the dictionary from the data itself. After a few iterations which allow for the algorithm to settle, we found that this choice leads to a better sparse encoding.

[4]It is also possible to take images in other similar categories for unlabeled data which we believe leads to no clear difference. Our choice here is to simplify and lead to a reproducible result.

[5]Note: it is of common practice in face recognition to crop a particular area of the face. For a pre-cropped version, this subset can be downloaded from [5]

Figure 3. Confusion matrix on Caltech 101.



for the other we discard the class information and use as *unlabeled* data. In this experiment, we use $\gamma = 1$ and $\rho = 1$. Other parameters are the same as used in the previous experiment.

For comparison, we also run the test against the regularized version of some recently proposed techniques in the face recognition literature [7][6] such linear discriminant analysis (LDA) [3], local preservation projection (LPP) [13] and orthogonal Laplacianfaces (OLPP) [6]. These regularized versions have been demonstrated to achieve very good classification performance even when the number of training samples is small [7]. In this experiment, we set the regularization parameter $\alpha = 0.1$ as suggested in [7] and use a simple quadratic regularizer$\mathcal{J}(\mathbf{a}) = \mathbf{a}^T\mathbf{a}$ (for the technical meaning please see Equation (11) in [7]).

To illustrate the advantage of our method over the compared methods when the number of training samples is small, we consider an extreme case of only 2 training samples and report the average classification error over all classes together with the standard deviation over 20 random splits. The result is tabulated in Table 1. As can be seen, the regularized versions R-LDA and R-LPP have outperformed the baseline PCA and OLPP as they are known to be quite robust in the small training size case. However, our proposed method clearly outperforms R-LDA by as much as 8%. When we increase the number of training samples, R-LDA and OLPP improve significantly and approximate our method at 4 training samples. Of course, there is some increase in the computational cost for our method, but this gain in performance justifies our proposed method.

3.3. Selection of the regularization parameter

As we mentioned, the parameter ρ in the formulation (5) controls the trade-off between representative error of both the labeled and unlabeled data and discriminative power.

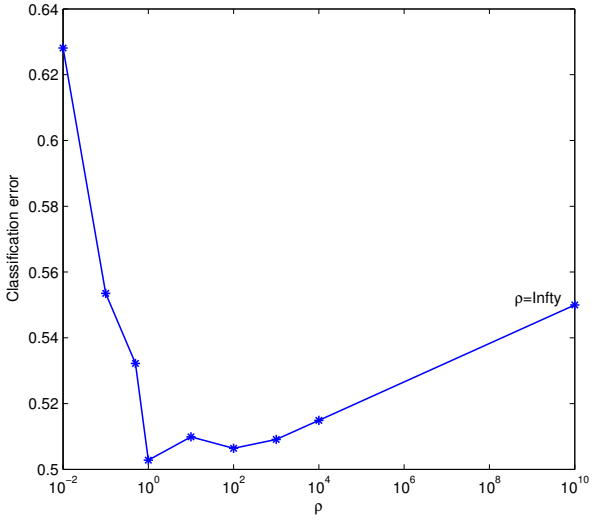
[6]The source code for these linear techniques is also obtained from [5]

Table 1. Error rate on Extended Yale B face database

Train	2	3	4
PCA	0.78 ± 0.02	0.76±0.02	0.74±0.02
R-LDA	0.58±0.02	0.43±0.03	0.36±0.03
R-LPP	0.62±0.03	0.54±0.03	0.50±0.02
OLPP	0.72±0.03	0.43±0.03	0.34±0.03
Proposed method	0.50±0.02	0.41±0.03	0.36 ± 0.02

By considering this trade-off in the formulation, it is believed that generalization ability will be better for classification than methods that are simply optimized for sparse representation such as [24], which is equivalent to setting ρ very large. A properly selected value for ρ can lead to better performance. Model selection methods such as cross-validation often examine the classification performance and select the value of ρ that is optimal. To address how the classification performance varies with different values of ρ, we revisit the face recognition experiment with varying ρ. The result is shown in Fig. 4. When $\rho = 0.01$, we lose much representative capability and the classification error jumps to about 62%. On the other hand, if we directly use the overcomplete sparse encoding coefficients directly from the K-SVD algorithm (i.e. equivalent to setting $\rho = \infty$) the classification error is 55%. Selecting a value of ρ in between these two extremes clearly leads to a better classification performance. However, future work needs to address a better selection of this parameter and an extension to the case where $\rho_l \neq \rho_u$.

Figure 4. Effect of varying ρ.



4. Conclusion

We have presented a new approach to pattern recognition using sparse overcomplete representations. The novelty of our proposed approach lies in the formulation of joint learning and dictionary construction that results in the most discriminative sparse encoding and an optimal linear classifier for the problem of interest. Our method solves this joint problem by formulating an optimization problem that involves both the objective function of the classification stage and the representation error of both the labeled and unlabeled data with sparsity constraint. A suboptimal algorithm has also been proposed to solve this constrained optimization problem. Being formulated in a statistical learning framework and exploiting the information in unlabeled data, our proposed method also achieves good generalization. Experimental results for object categorization over the Caltech 101 database show that with a very simple classifier, our proposed method is already competitive with many recently proposed alternatives. When being applied to face recognition, our method clearly outperforms recently proposed robust methods when the number of training samples is small. These early results encourage further exploration in this framework both theoretically and practically.

Acknowledgement

We thank the anonymous reviewers for helpful comments that improve the clarity of the paper. This work is in part supported by the Australian Research Council.

Appendix I

The result of Lemma 1 is similar to the Power method in finding the maximum eigenvalue. Here we seek a rank-1 approximation $\mathbf{a}\mathbf{b}^T$ of a (not necessarily square) matrix $\mathbf{M} \in \mathbb{R}^{p\times q}$. When both $\mathbf{a}$ and $\mathbf{b}$ are free, the standard SVD algorithm yields the optimal solution in a sense of minimising the squared distance. However, when there are constrainst on $\mathbf{a}$ or $\mathbf{b}$, Lemma 1 serves as a suboptimal way to find the solution.

To prove (18), denote $\mathbf{c}_i$ the ith column of $\mathbf{M}$, then a bit of algebra gives

$$\begin{aligned} f(\mathbf{b}) &= \| \mathbf{M} - \mathbf{a}\mathbf{b}^T \|_F^2 \\ &= \sum_{i=1}^{q} \| \mathbf{c}_i - \mathbf{a}b_i \|^2 . \end{aligned} \tag{27}$$

Minimising $f(\mathbf{b})$ for $\mathbf{b} \in \mathbb{R}^q$ yeilds $b_i = \mathbf{c}_i^T\mathbf{a}/ \| \mathbf{a} \|^2$, or equivalently

$$\mathbf{b} = \frac{\mathbf{M}^T\mathbf{a}}{\| \mathbf{a} \|^2}. \tag{28}$$

To prove (19), we can expand the above expression

$$f(\mathbf{a}) \quad = \quad \sum_{i=1}^{q}(\mathbf{c}_i^T\mathbf{c}_i - 2b_i\mathbf{c}_i^T\mathbf{a} + \mathbf{a}^T\mathbf{a}). \qquad (29)$$

From which minimising $f(\mathbf{a})$ subject to $\mathbf{a}^T\mathbf{a} = 1$ is equivalent to

$$\min_{\mathbf{a}^T\mathbf{a}=1} \sum_{i=1}^{q} -2b_i\mathbf{c}_i^T\mathbf{a} = \min_{\mathbf{a}^T\mathbf{a}=1} -2\mathbf{a}^T\mathbf{M}\mathbf{b}. \qquad (30)$$

The solution for $\mathbf{a}$ is then the unit norm vector in the direction of $\mathbf{Mb}$ which is

$$\mathbf{a} = \frac{\mathbf{Mb}}{\parallel \mathbf{Mb} \parallel}. \qquad (31)$$

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*, 54:4311–4322, Nov 2006.

[2] S. An, W. Liu, and S. Venkatesh. Face recognition using kernel ridge regression. In *Proceedings CVPR*, 2007.

[3] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

[4] O. Bryt and M. Elad. Compression of facial images using the K-SVD algorithm. *IEEE Transactions on Image Processing*, (under review).

[5] D. Cai. Codes and datasets for subspace learning. Available online at www.cs.uiuc.edu/homes/dengcai2/Data/data.html, As retrieved on October 2007.

[6] D. Cai, X. He, J. Han, and H. J. Zhang. Orthogonal Laplacianfaces for Face Recognition. *IEEE Transactions on Image Processing*, 5(11):3608–3614, 2006.

[7] D. Cai, X. He, Y. Hu, J. Han, and T. Huang. Learning a spatially smooth subspace for face recognition. In *Proceedings CVPR*, 2007.

[8] D. L. Donoho, M. Elad, and V. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory*, 29(1):6–18, Jan 2007.

[9] M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 15(12):3736–3745, Dec 2006.

[10] L. FeiFei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental-bayesian approach tested on 101 object categories. In *Proceedings CVPR workshop on generative model based vision*, 2004.

[11] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):643–660, 2001.

[12] F. Girosi. An equivalence between sparse approximation and support vector machines. *Neural computation*, 10(6):1455–1480, 1998.

[13] X. He and P. Niyogi. Locality preserving projection. In *Proceedings NIPS*, 2003.

[14] A. E. Hoerl and R. W. Kennard. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12:105–123, 1970.

[15] K. Huang and S. Aviyente. Sparse representation for signal classification. In *NIPS'06*, 2006.

[16] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky. A method for large-scale ℓ_1-regularized least squares. *IEEE Journal on Selected Topics in Signal Processing*, to appear. (Software package available at www.stanford.edu/~boyd/l1_ls/).

[17] Y. LeCun, M. Ranzato, and F.-J. Huang. Energy-based models in document recognition and computer vision. In *Proceedings of the International Conference on Document Analysis and Recognition(ICDAR)*, 2007.

[18] M. S. Lewicki and T. J. Sejnowski. Learning overcomplete representations. *Neural computation*, 12:337–365, 2000.

[19] J. Mairal, M. Elad, and G. Sapiro. Sparse representation for color image restoration. *IEEE Transactions on Image Processing*, (to appear).

[20] J. Mairal, G. Spairo, and M. Elad. Multiscale sparse image representation with learned dictionaries. In *Proceedings of the International Conference on Image Processing (ICIP)*, San-Antonio Texas, September 16-19 2007.

[21] J. Mutch and D. G. Lowe. Multiclass object recognition with sparse, localized features. In *Proceedings CVPR*, 2006.

[22] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive-field properties by learning sparse codefornatural images. *Nature*, 381:607–609, 1996.

[23] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employ by v1? *Vision Research*, 37:3311–3325, 1997.

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

[25] M. Ranzato, F. J. Huang, Y.-L. Boureau, and Y. LeCun. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *Proceedings CVPR*, 2007.

[26] P. Salle and B. A. Olshausen. Learning sparse multiscale image representations. In *Proceedings NIPS*, 2003.

[27] T. Serre, L. Wolf, S. Bielschi, M. Riesenhuber, and T. Poggio. Robust object recognition with cortex-like mechanism. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):411–426, Mar 2007.

[28] M. V. Srivastava. *Methods of multivariate statistics*. Wiley-Interscience, 2002.

[29] J. A. Tropp. Just relax: convex programming methods for identifying sparse signals innoise. *IEEE Transactions on Information Theory*, 52:1030, Mar 2006.

[30] V. Vapnik. *Statistical learning theory*. Wiley, 1998.

[31] M. Varma and D. Ray. Learning the discriminative power-invariance trade-off. In *Proceedings ICCV*, 2007.