Discriminative sparse model and dictionary learning 
for object category recognition

Xiao Deng and Donghui Wang*
Institute of Artificial Intelligence, Zhejiang University 
Hangzhou, China, 310027
{yellowxiao,dhwang}@zju.edu.cn

Abstract—Recent researches have well established that sparse signal models have led to outstanding performances in signal, image and video processing tasks. This success is mainly due to the fact that natural signals such as images admit sparse representations of some redundant basis, also called dictionary. This paper focuses on learning discriminative dictionaries instead of reconstructive ones. It has been shown that discriminative dictionaries, which are composed of sparse reconstruction and class discrimination terms, outperform reconstructive ones for image classification tasks. Experimental results in image classification tasks using examples from the Caltech 101 Object Categories show that the proposed method is efficient and can achieve a higher recognition rate than reconstructive methods.

Keywords—Sparse representation, dictionary learning, object category recognition, classification

I. INTRODUCTION

Category-level object recognition deals with one kind of visual tasks where all instances of a category are to be recognized. The key challenge in this problem consists in how to learn representative models of visual object categories for recognition. The learned object models are required not only to capture common semantic information of certain category, but also to generate exactly discriminative capability. In this paper, we focus on this challenging problem and adapt a discriminative sparse model and dictionary learning method for object category recognition.

An important representative model for category-level object recognition is the bag of words (BoW) approach which has been investigated by several studies[1], [2], [3]. The key idea behind this model is to learn a visual codebook (SIFT descriptors) for each category by using an unsupervised clustering algorithm such as k-means. One major problem with this approach is that the construction of visual codebook and classifier training are two disconnected stage. That says the visual codebook of a category is learned separately without using the information of other categories. This reduces the discriminative ability of learned codebooks.

Another competitive model for category-level object recognition is sparse signal representation[4]. In this model, the sparsity is confirmed as a powerful prior for visual inference and the semantic information of the image is naturally encoded as sparse representation with respect to predefined bases or learned dictionaries. The key problem in this model is how to correctly choose the basis or dictionary for representing image information. Using predefined bases such as curvelets, wedgelets, wavelets and its variants[12], isn’t suitable for specific vision tasks because these bases don’t use the semantic prior of given sample images. Learning a task-specific dictionaries instead can lead to more significant improvements than using predefined ones[4].

In current sparse representation and dictionary learning methods, the main goal of learning a dictionary for sparse representation is to reconstruct image information instead of category recognition. To recognize an object category needs to combine sparse reconstruction and class discrimination. This motivates us to learn a discriminative sparse model for representing common semantic information of specific category as well as for discriminating different object categories. To achieve this purpose, we follow the strategy proposed in[5] for local image analysis. $N$ dictionaries for $N$ categories are learned simultaneously under the constraint of reconstructive and discriminative loss function. The reconstruction errors of learned dictionaries on image patches are used to derive an object category recognition.

The paper is organized as follows. Section 2 presents background of sparse representation and dictionary learning. A discriminative sparse model and dictionary learning method for object category recognition is introduced in Section 3. Section 4 presents experimental results on object categories recognition. Concluding remarks are given in Section 5.

II. BACKGROUND AND RELATED WORK

In classical sparse coding problem, one considers a signal $y \in \mathbb{R}^n$ and a dictionary $D = [d_1, ..., d_k] \in \mathbb{R}^{n \times k}$. Under the assumption that a signal can be approximately represented as a sparse linear combination of the atoms from the dictionary, $y$ can be approximately represented as $D\alpha$ for some sparse coefficient vector $\alpha \in \mathbb{R}^k$. Thus finding the sparse representation of $y$ leads to the following optimization problem:

$$
\min_{\alpha \in \mathbb{R}^k} ||\alpha||_p, \text{ s.t. } ||y - D\alpha||_2 \leq \varepsilon
$$

(1)

where the most frequently-used values for $p$ are 0 and 1. However, when $p$ is 0, the solution of this $\ell^0$ norm optimization problem has been proven to be an NP-hard problem. Thus, approximate solutions are considered, and several algorithms such as Matching pursuit (MP)[7] and Orthogonal Matching pursuit (OMP)[6] have been proposed in the past decades. When $p$ is 1, this $\ell^1$ norm optimization problem is the

*Correspondence author.
well-known Basic Pursuit[9] or Lasso[8] problem. Several algorithms such as LARS[10] have been proposed to solve this problem efficiently.

Recent study has demonstrated that using learned dictionaries rather than predefined ones can lead to state-of-the-art performances[4]. In practice, overlapping patches instead of the whole images are sparsely decomposed because the natural images are usually very large. For an image $I$, suppose there are $M$ overlapping patches $\{y_j\}_{i=1}^M \in \mathbb{R}^n$ of size $n$ from the image $I$. Then the dictionary $D \in \mathbb{R}^{n \times k}$ is learning via solving the following optimization problem:

$$
\min_{D, \alpha} \sum_{i=1}^{M} \|\alpha_i\|_p, \ \text{s.t.} \ \|y_i - D\alpha_i\|_2^2 \leq \epsilon \tag{2}
$$

where $A = [\alpha_1, ..., \alpha_M]$ is a matrix in $\mathbb{R}^{k \times M}$, $y_i$ is the $i$-th patch of the image $I$ written as a column vector, $\alpha_i$ is the corresponding sparse code, $D\alpha_i$ is the estimate of $y_i$ and the dictionary $D$ is normalized with each column having an unit $\ell^2$ norm. Several algorithms have been proposed for addressing the above optimization problem (2). Among these algorithms, the K-SVD algorithm has been widely used for reconstruction tasks in recent years[11].

III. DISCRIMINATIVE SPARSE MODEL

In our task, some sample images are provided as training set. They are classified in advance by object categories. We need to learn the representative models of object categories as prior knowledge for recognizing the category of test image. As introduced above, the current sparse model provides an effective framework to extract common and compact representation of one category by dictionary learning algorithm. Such dictionaries are learned separately for each category and obtain optimal decomposition basis for image reconstruction. For object category recognition, the reconstruction residual of test image on learned dictionary of each category is used as a criteria for discrimination. The flow of this method is shown in Figure 1. It’s worth notice that dictionary learning of one category don’t use any information of other categories which may be an important cue for discrimination. Thus, we present a discriminative sparse model and dictionary learning algorithm to seek optimal representative basis for object category recognition.

Consider $C$ sets of signals $\{(y_1)_{i=1}^n\}_{j=1}^C \in \mathbb{R}^{n \times N}$, where $N$ denotes the total number of signals. Our goal is to learn $C$ discriminative dictionaries $\{D_j\}_{j=1}^C \in \mathbb{R}^{n \times kC}$, making dictionary $D_j$ corresponding to the $j$-th class good at reconstructing this class while bad for the other classes simultaneously. Following the strategy in [5], we make some improvements. In our framework, the residual $\|y - D\alpha\|_2^2$ is used as a criteria for discrimination. To achieve the goal of discrimination, we introduce a discriminative function, which is the log function of the reciprocal of the softmax function

$$
f_j(x) = C_j(x_1, ..., x_C) = \log(\sum_{i=1}^{C} e^{(x_i - x_j)}) \tag{3}
$$

Thus, finding the class label of testing sets is equivalent to minimize (3). Then, adding this discriminative term into the original dictionary updating formulation, we have

$$
\min_{\{D_j\}_{j=1}^C} \sum_{i=1}^{n} \left( \sum_{l=1}^{C} C_l(\|y_l - D_j\alpha_l\|_2^2) + \gamma\|y_l - D_i\alpha_i\|_2^2 \right) \tag{4}
$$

where $\gamma$ is the parameter governing the trade-off between reconstruction and discrimination, and $\alpha_j$ denotes the corresponding sparse coefficients of signal $y_l$ on the dictionary $D_j$ which have been computed during the sparse coding step. Solving this optimization problem is hard because it is neither convex nor differentiable. Thus, truncated Newton method is considered to give an approximate formulation easier to solve. Then, updating the $p$-th dictionary is equivalent to solve the following formulation

$$
\min_{D \in \mathbb{R}^{n \times k}} \sum_{i=1}^{n} \left( \sum_{l=1}^{C} w_l \|y_l - D\alpha_{lp}\|_2^2 \right), \ \text{where}
$$

$$
w_l = \frac{\partial C_l}{\partial x_l}(\{\|y_l - D\alpha_{lp}\|_2^2\} + \gamma \mathbf{1}_p(i)) \tag{5}
$$

$x_p$ denotes $\|y_l - D_p\alpha_{lp}\|_2^2$, $\mathbf{1}_p(i)$ is 1 if $i = p$ and 0 otherwise. Inspired by the idea of the K-SVD algorithm, (5) can be rewritten as

$$
\min_{d_j, \alpha} \sum_{i=1}^{C} \sum_{l \in \{1, ..., n\} \cap \omega_l} w_l \left\| (y_l - d_j - \alpha_{lp}) - d_j \alpha_p \right\|_2^2 \tag{6}
$$

where $\omega_l = \{i | 1 \leq i \leq N, \text{s.t} \ \alpha'(i) \neq 0\}$ denotes indices of signals that use the atom $d_j$, $d_p$ and $\alpha_p$ denote all the $d_j$ and $\alpha_j$ whose indices are not $p$. The matrix form of the above formulation is

$$
\min_{d_j, \alpha} \left\| (Y - DA)W^\frac{1}{2} \right\|_F = \min_{d_j, \alpha} \left\| (E_j - d_j \alpha_j^*)W^\frac{1}{2} \right\|_F \tag{7}
$$

where $Y = [y_1, ..., y_N] \in \mathbb{R}^{n \times N}$ is the signal matrix, $A = [\alpha_1, ..., \alpha_N] \in \mathbb{R}^{k \times N}$ is the coefficient matrix, and $W = \text{diag}(w_1, ..., w_N)$ is the diagonal matrix of the weights $w_l$. Using the definition of condensed SVD decomposition, matrix $E_jW^{\frac{1}{2}}$ can be written as $U\Sigma V^T$. Then, the solution for atom $d_j$ is $u_1$, which is the eigenvector associated to the largest eigenvalue of $E_jW^{\frac{1}{2}}$, and the solution for the non-zero coefficient $\alpha_j$ is $((W^{\frac{1}{2}})^{-1}\lambda_1v_1)^T$. The details are shown in Algorithm 1.

Suppose we have obtained $C$ dictionaries $D_j$ one for each class, via the above discriminative K-SVD algorithm. The strategy for classification in our paper is based on the reconstructive errors, which are generated by approximating each patch using the $C$ dictionaries $D_j$. Thus, finding the right class label of patch $y$ is equivalent to solve

$$
\min_i \|y - D_i\alpha_i\|_2^2 \tag{8}
$$

IV. EXPERIMENTAL RESULTS

In this section, we study whether our proposed discriminative framework can enhance the discriminative power of dictionaries and is more efficient when applied to object category recognition tasks. Meanwhile, we also study the roles
of the three parameters in our proposed framework size of patches, size of dictionaries, and $\gamma$ via our experiments.

We have selected 50,000 patches which are centered and have normalized $\ell^2$ norm, respectively from images of different classes from the Caltech 101 Object Categories dataset. There are two different sizes of patches which are $20 \times 20$ and $32 \times 32$ in our experiments. Figure 2 shows the partial original patches of these two different sizes. We then divide patches of each category into a training set and a testing set, respectively of 15,000 and 35,000 patches.

### A. Choices of parameters

**Size of patches.** In this experiment, we study the role of size of patches in our proposed framework. We use two groups of patches of size $20 \times 20$ and $32 \times 32$ from natural images of bike and background. The size of dictionaries $k$ is 144, the sparsity factor $L$ is 4, and $\gamma$ is 3.0. Then 15 iterations of our framework are processed. Figure 3 shows the dictionaries learned from the two groups of patches of the two sizes. The error rates of these two groups of dictionaries are showed in Table I.

Comparing the dictionaries of bike with those of background, we can see that there are more arc-shaped patches in the dictionaries of bike while in the dictionaries of background there are more bar-shaped patches. Those arc-shaped patches correspond to the wheels of bikes while in the images of background there are rarely things of round or other shapes. Meanwhile, comparing the dictionaries of patches of two sizes, we observe that the dictionary of bike learned from patches of larger size has more arc-shaped patches than that learned from patches of smaller size and dictionaries of larger patches are more distinguishable than those of smaller patches. Moreover, from the error rates showed in Table I we observe that dictionaries of larger patches are more discriminative than those of smaller patches. Patches of small size contain more reconstructive information, however, lose important information that distinguishes one category from others. Thus, when learning dictionaries applied for discrimination tasks, patches of large size are selected.

**Size of dictionaries.** This experiment studies the role of size of dictionaries in our framework. We use patches of size $32 \times 32$ from natural images of bike and background, but this time we change the size of dictionaries $k$ into 64 and $\gamma$ remains unchanged. Figure 3 shows the dictionaries learned in this experiment and the error rates of this experiment are

### Table I

<table>
<thead>
<tr>
<th>Class</th>
<th>$p = 400, k = 144$</th>
<th>$p = 1024, k = 144$</th>
<th>$p = 1024, k = 64$</th>
</tr>
</thead>
<tbody>
<tr>
<td>bike</td>
<td>7.5%</td>
<td>2.5%</td>
<td>5%</td>
</tr>
<tr>
<td>none</td>
<td>2.5%</td>
<td>1%</td>
<td>2.5%</td>
</tr>
</tbody>
</table>

### Diagram

(a) Dictionary learning stage.

(b) Object category recognition stage.

![Diagram](image-url)
Algorithm 1 The Algorithm of discriminative sparse model

Input: $N$ signals $\mathbf{y}_l \in \mathbb{R}^n$ which belong to $C$ different classes.
Output: $C$ dictionaries $D_l \in \mathbb{R}^{n \times k}$.

Initialization: Initialize the $C$ dictionaries $D_l$ with $\ell^2$ normalized columns.

Loop: Repeat until convergence (or stopping criterion):
1. Sparse coding: For $C$ dictionaries $D_l$ and $N$ vectors $\mathbf{y}_l$ finding the solution of the following optimization problem:
   $$
   \min_{\mathbf{a}_l} \|y_l - D_l \mathbf{a}_l\|_2^2, \quad \text{s.t.} \quad \|a_l\|_p \leq L
   $$
2. Dictionary updating: For $i = 1, 2, ..., C$, for $j = 1, 2, ..., k$, update $d_j$, the $j$-th atom of $D_l$:
   - Select the set of signals using atom $d_j$:
     $$
     \omega_j = \{l \in 1, 2, ..., N | a_l[j] \neq 0\}
     $$
   - If there are no indices in set $\omega_j$:
     - For $l = 1, 2, ..., N$, find the least represented signal $y_l$ via solving the following maximization problem:
       $$
       y_l = \arg \max_{y_l} \|y_l - D_l \mathbf{a}_l\|_2^2
       $$
     - Update the $j$-th atom $d_j$ using the normalized signal $y_l$, and the corresponding non-zero coefficients using zero vector.
   - Else:
     - Compute the representation error matrix $E$:
       $$
       E = Y - \sum_{p \neq j} d_p \alpha_l^p
       $$
     - Restrict $E$ to the columns included in the set $\omega_j$, and obtain $E_j$.
     - Compute the weights $w_l$ in $\omega_j$.
     - Update a new $d_j'$ and the associated nonzero coefficients $\alpha_{l_i}'$ via solving the following optimization problem:
       $$
       \min_{d_j', \alpha_{l_i}'} \left\| \left( E_j - d_j \alpha_l \right) W_j^{\frac{1}{2}} \right\|_F^2
       $$
   - The new $d_j'$ is the eigenvector associated to the largest eigenvalue of $E_jW_j^{\frac{1}{2}}$, $W = \text{diag}(w_1, ..., w_N)$.

showed in Table I.

Comparing dictionaries learned in this experiment with those of size 144 which are learned under the same conditions showed in Figure 3, we observe that dictionaries learned in the last group of experiments are more distinguishable from each other than those learned this time. Meanwhile, from the results showed in Table I we can see that the discrimination power of dictionaries of smaller size is weaker than that of dictionaries of larger size. If the size of dictionaries is too small, information in learned dictionaries is not sufficient for discrimination. Thus, the size of dictionaries should be large enough for helping dictionaries more discriminative when applied them for object category recognition tasks.

Parameters $\gamma$. In this group of experiments, we study the role of parameters $\gamma$ in our framework. Meanwhile, explore the role of $\lambda$ introduced in Mairal’s paper[5]. We use patches of larger size in the first group of experiments, and the other parameters remain the same as our first group of experiments except the values of parameters $\gamma$ and $\lambda$. In these experiments, we use different values of parameters $\gamma = \{1, 1.5, 3, 5, 7.5, 10, 12, 15, 20, 50\}$, and $\lambda = \{0.1, 0.2, 0.5, 0.75, 1, 1.5, 2, 3, 8, 10\}$. Figure 4 shows the residuals of the training patches on learned dictionaries of bike and background. The percentage error rates when $\gamma = 1$ for any $\lambda$ of bike and background are $\{62.5, 25, 57.5, 77.5, 50, 55, 27.5, 30, 60, 47.5\}$ and $\{35, 57.5, 60, 57.5, 67.5, 72.5, 67.5, 52.5, 52.5, 67.5\}$ and the error rates of other $\gamma$ and $\lambda$ are close and low.

From Figure 4, we observe that when the values of parameter $\gamma$ are larger than some value, the residuals remain roughly the same no matter what the values of $\gamma$ and $\lambda$ are. Thus, we omit parameter $\lambda$ in our framework. The reconstructive parameter $\gamma$ controls the trade-off between reconstruction and discrimination. (4) has solution if and only if the diagonal matrix $W$ of the weights $w_l$ is positive definite, and from (5) we can easily see that $\gamma$ plays an important role computing the weights $w_l$. Thus, the value of $\gamma$ should be large enough to ensure the existence of the solution of our proposed discriminative formulation. Meanwhile, if $\gamma$ is too large, the discriminative power will reduce according to (4). Thus, the value of $\gamma$ should small enough to ensure the discriminative ability.

B. Object category recognition

This subsection studies whether our proposed framework is more efficient than other methods of object categories recognition and also efficient for multiclass recognition tasks. In our first group of experiments, we have mainly compared our method with the original reconstructive dictionary learning and the BoW methods. We use the patches of size $20 \times 20$, and the size of dictionaries is 64. We want to study whether our method has an advantage over the reconstructive one when the parameters are good for reconstruction while bad for discrimination. 15 iterations are performed in our framework and the classical dictionary learning framework. Figure 5 shows the dictionaries learned via these two methods. The discrimination results are showed in Table II. From results...
showed in Table II, we can easily see that the discrimination power of our method is more efficient than that of the classical dictionary learning and BoW methods.

This group of experiments shows that our proposed framework is also efficient when applied to multiclass recognition tasks. We use patches of size $32 \times 32$ from natural images of butterfly, ketch, and face. The size of dictionaries $k$ is 144, the sparsity factor $L$ is 4, and $\gamma$ is 3. 20 iterations of our framework are performed. Figure 6 shows the partial natural images used in this experiment. Figure 7 shows the dictionaries learned in this experiment and Table III shows the error rates of these three classes.

<table>
<thead>
<tr>
<th>Image class</th>
<th>discrimination</th>
<th>reconstruction</th>
<th>BoW</th>
</tr>
</thead>
<tbody>
<tr>
<td>bike</td>
<td>5%</td>
<td>12.5%</td>
<td>7.5%</td>
</tr>
<tr>
<td>background</td>
<td>1%</td>
<td>7.5%</td>
<td>2.5%</td>
</tr>
</tbody>
</table>

Fig. 6. Partial natural images used in this experiment.

V. CONCLUSION

This paper proposed a discriminative sparse models and dictionary learning method for object category recognition. Experiments with the Caltech 101 Object Categories show that our framework is efficient when applied for object category recognition tasks. Meanwhile, we also find out the roles of the four parameters size of patches, size of dictionaries, and $\gamma$ via comparing these experiments.

ACKNOWLEDGMENTS

This work is supported by 973 Program (Project No. 2010CB327905) and Natural Science Foundations of China (No.61071218).

REFERENCES