JMLR: Workshop and Conference Proceedings 25:491–505, 2012 Asian Conference on Machine Learning

Spatial Locality-Aware Sparse Coding and Dictionary Learning

Jiang Wang JWA368@EECS.NORTHWESTERN.EDU
2145 Sheridan Road, Evanston IL 60208

Junsong Yuan JSYUAN@NTU.EDU.SG
School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798

Zhuoyuan Chen ZCH318@EECS.NORTHWESTERN.EDU
2145 Sheridan Road, Evanston IL 60208

Ying Wu YINGWU@EECS.NORTHWESTERN.EDU
2145 Sheridan Road, Evanston IL 60208

Editor: Steven C.H. Hoi and Wray Buntine

Abstract

Nonlinear encoding of SIFT features has recently shown good promise in image classification. This scheme is able to reduce the training complexity of the traditional bag-of-feature approaches while achieving better performance. As a result, it is suitable for large-scale image classification applications. However, existing nonlinear encoding methods do not explicitly consider the spatial relationship when encoding the local features, but merely leaving the spatial information used at a later stage, e.g. through the spatial pyramid matching, is largely inadequate. In this paper, we propose a joint sparse coding and dictionary learning scheme that take the spatial information into consideration in encoding. Our experiments on synthetic data and benchmark data demonstrate that the proposed scheme can learn a better dictionary and achieve higher classification accuracy.

1. Introduction

Image classification is a challenging problem for computer vision. Bag-of-Feature (BOF) approaches Fergus et al. (2003) summarize local features by quantizing each feature into a "visual word" and represent an image by a histogram of the "visual words". These approaches achieve good results because they are relatively robust to the intra-class variations. However, BOF approaches largely discard the spatial relationship among the local features. Spatial Pyramid Matching (SPM) Lazebnik et al. (2006) approach is among the most successful methods so far that takes the spatial layout into consideration. Motivated by Grauman and Darrell (2005), the SPM approach partitions the image into increasingly finer spatial sub-regions and computes the histogram of local features for each sub-region. The resulting spatial pyramid achieves a good balance between the incorporation of the spatial information and the robustness to the intra-class variations. It is computationally efficient, and has shown very good performance on many image classification tasks. A schematic framework of the SPM approach is illustrated in Fig. 1(a).

However, many experiments indicated that the SPM approach has to use a nonlinear kernel SVM to achieve good performance. The nonlinear classifier is very expensive to train in practice. Training a nonlinear SVM has a complexity of $O(n^3)$, where n is the number of support vectors.

In order to improve scalability, researchers have attempted to find some nonlinear features that work better with linear classifiers, which has an $O(n)$ computational complexity in training. Specif-

ically, Yang et al. (2009) proposed the ScSPM approach, whose framework is illustrated in Fig. 1(b). ScSPM approaches employ sparse coding to encode the SIFT features into sparse vectors, and use max-pooling to pool the sparse vectors in each sub-region to form a spatial pyramid. It has been shown that even with a linear SVM, the ScSPM approach can achieves good performance in many benchmark datasets.

However, the ScSPM approach does not explicitly consider the spatial information in sparse coding. This information can be very helpful for both sparse coding and dictionary learning. In this paper, we propose a sparse coding scheme called *spatially locality-aware sparse coding* that takes the local sparsity of the features into account. We depict this new method in Fig. 1(c). The proposed sparse coding scheme is based on the following observation. For a local feature, its spatial context in general only consists of a limited set of different features in the neighborhood. These features in its spatial context are from a small fraction of the dictionary. Therefore, it is more plausible that the similar features bearing similar spatial contexts should be encoded as the same sparse vector. This implies that the spatial context of a feature should also be sparse. We call this the *spatially local sparsity* property. This property can be used for better coding and dictionary learning. This is different from traditional sparse coding methods. Exploiting this information in sparse coding, we are able to obtain more stable features and learn a better dictionary. The proposed sparse coding approach has been tested in a synthetic dataset and the Caltech-101 dataset, and achieved promising results.

The contributions of this paper are as follows. First, we propose a novel sparse coding and dictionary learning approach for image classification that exploit spatial locality information by regularizing the sparseness of the pooled vector in each subregion. Secondly, an efficient optimization algorithm based on FISTA Beck and Teboulle (2009) is derived to solve the corresponding optimization problem.

The rest of the paper is organized as follows. Section 2 reviews the related work. Section 3 briefly summarizes the sparse coding spatial pyramid approach for image classification. Section 4 introduces the proposed spatial locality-aware sparse coding approach and derives an efficient algorithm to solve the corresponding optimization problem. Section 5 describes an efficient online dictionary learning algorithm. Finally, in Section 7, we conclude our paper, and discuss some future research directions.

2. Related Work

In recent years, sparse coding and related techniques have been successfully applied to object recognition applications. In addition to the work by Yang et al. (2009), various coding schemes have been proposed recently. Wang et al. (2010) proposes a locality-constrained coding scheme to efficiently encode the local features. Yu and Zhang (2010); Yu et al. (2009) employ local vectors and their tangents on data manifold to encode the local features. In Zhou et al. (2010); Gao et al. (2010a), local features are encoded by utilizing a nonlinear kernel. Gao et al. (2010b) encourages similar features to have similar sparse vectors by adding a Laplacian distance matrix regularization, and Yang et al. (2010b) use a mixture of overcomplete dictionaries to enable the usage of a large dictionary in sparse coding. Sparse coding was also used to encode Gabor filter response features for palmprint verification Zuo et al. (2010). However, none of those work considers the spatial information explicitly in feature encoding.

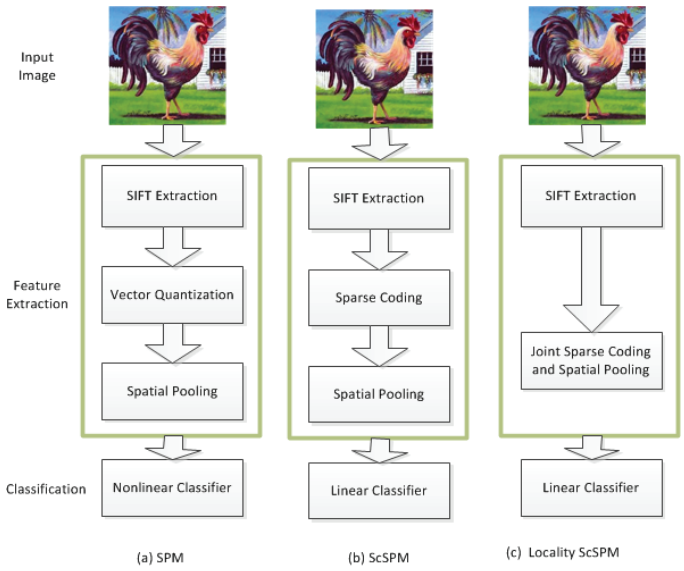


Figure 1: Schematic comparison of the original SPM, the SPM based on sparse coding (ScSPM), and the proposed spatial locality-aware sparse coding

Dictionary learning, also known as codebook optimization, is also a very active research area. Lee et al. (2007) proposes to use a coordinate descent method for the optimization and to update the dictionary by solving the Lagrangian dual of the optimization problem. In Mairal et al. (2009), an efficient online learning algorithm for dictionary learning is proposed. Supervised dictionary learning has also been studied. In Yang et al. (2010a), a back-propagation approach is used to update the dictionary using the labeled data. Ramirez et al. (2010) learn incoherent and shared dictionary for each class. By taking the spatial information explicitly into account, our proposed method is able to learn a better dictionary for classification tasks.

There are some other works that are also related to our paper. Yuan and Wu (2008) try to optimize the clustering codebook by jointly clustering the local contextual features and the local features.

3. Image Classification with Sparse Coding

To make this paper self-contained, this section summarizes the image classification scheme based on sparse coding Yang et al. (2009). The framework of image classification with sparse coding is illustrated in 1(b). First, local descriptors are extracted from all images. SIFT features are usually utilized as local descriptors. Given these descriptors and a dictionary, sparse coding computes a sparse vector for each descriptor. Then, max-pooling can be employed to summarize the sparse vectors in each region hierarchically into a spatial pyramid. Finally, a linear SVM can be used to train a discriminative model for image classification. Some details of this framework are as follows.

Given a set of M-dimensional local descriptors extracted from images, $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N] \in \mathbb{R}^{M \times N}$, and given a codebook with D bases:$\boldsymbol{B} = [\boldsymbol{b}_1, \boldsymbol{b}_2, \cdots, \boldsymbol{b}_D] \in \mathbb{R}^{M \times D}$, sparse coding aims

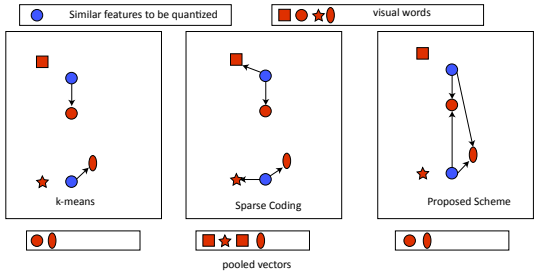


Figure 2: Comparison of different encoding schemes. In k-means, each feature is only assigned to one clustering center; in sparse coding, features are automatically assigned to the centers that can optimally reconstruct this feature. In the proposed scheme, by constraining the sparsity of the pooled vector, the similar features in the same spatial context are encouraged to have similar sparse vector. As a result, it can produce more stable results.

to convert each descriptor into a D-dimensional code to obtain the final image representation by optimizing the following objective function.

$$\min_{\boldsymbol{Z}} \sum_{i=1}^{N} \|\boldsymbol{x}_i - \boldsymbol{B}\boldsymbol{z}_i\|_2^2 + \lambda \sum_{i=1}^{N} \|\boldsymbol{z}_i\|_1 \tag{1}$$

where $\boldsymbol{Z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots \boldsymbol{z}_N] \in \mathbb{R}^{D \times N}$ are the sparse vectors.

The sparse vectors can be considered as soft vector quantization of the features. Compared to hard vector quantization, soft vector quantization is more stable, because hard vector quantization, which usually takes the class label of the nearest basis, may be very different for similar descriptors, as illustrated in Fig. 2. Thus soft quantization often achieve better performance in image classification tasks Wang et al. (2010).

Max-pooling has been shown to be a good approach to combine multiple sparse vectors in a sub-region Yang et al. (2009). In each subregion, the max-pooling operation is defined as a function of the sparse vectors $\boldsymbol{z}_1, \cdots, \boldsymbol{z}_{N_L}$ in this subregion

$$\boldsymbol{\beta} = \xi_{\max}(\boldsymbol{z}_1, \cdots, \boldsymbol{z}_N), \tag{2}$$

where $\xi_{\max}$ is defined on each elements of $\boldsymbol{z}_i$. The i-th element of $\boldsymbol{\beta}$ is the maximum of the absolute values of the i-th element of all $\boldsymbol{z}_i$, i.e.,

$$\beta_j = \max\{|z_{j1}|, |z_{j2}|, \cdots, |z_{jN}|\} \tag{3}$$

where β_j is the j-th element of $\boldsymbol{\beta}$, and z_{ji} is the j-th element of $\boldsymbol{z}_i$. An illustration of max-pooling is shown in Fig. 3.

By hierarchically pooling the feature vector of each subregion, we can form a spatial pyramid. The features vectors in this spatial pyramid are concatenated to form the feature vectors of the whole image.

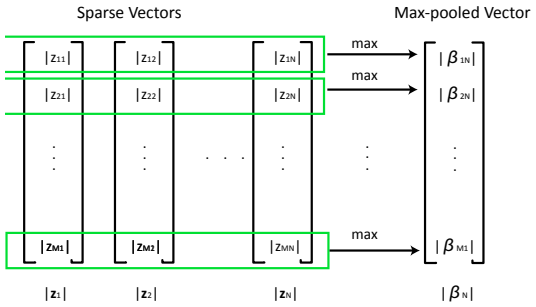


Figure 3: The illustration of the maximum pooling

4. Spatial Locality-Aware Sparse Coding

The sparse coding method minimizes the objective function in Eq. (1) for each local descriptor separately. In this section, we propose a novel sparse coding scheme called *spatial locality-aware sparse coding* that exploits the spatially local sparsity of the local descriptors. By exploiting the local sparsity constraint of local features, we are able to obtain more stable features, because similar features that in the same local contexts are encouraged to be encoded as similar sparse vectors under this constraint. This is illustrated in Fig. 2.

4.1. Formulation

In order to exploit the spatial local sparsity property of the local features, in our formulation, instead of regularizing individual encoded feature vectors to be sparse, we regularize the representation of the locally pooled region to be sparse.

In each local region, given a set of M-dimensional local descriptors extracted from this region, $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N]\mathbb{R}^{M\times N}$, and a codebook with D bases:$\boldsymbol{B} = [\boldsymbol{b}_1, \boldsymbol{b}_2, \cdots, \boldsymbol{b}_D] \in \mathbb{R}^{M\times D}$, the regularized objective function is

$$\min_{\boldsymbol{Z}} \sum_{i=1}^{N} \|\boldsymbol{x}_i - \boldsymbol{B}\boldsymbol{z}_i\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1, \tag{4}$$

where $\boldsymbol{Z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots \boldsymbol{z}_N] \in \mathbb{R}^{D\times N}$ are the sparse vectors, and $\boldsymbol{\beta}$ is the region level feature over the sparse vectors $\boldsymbol{z}_i$ by the max pooling defined in Eq. (2).

Similar hierarchical spatial pooling can be applied to the sparse vectors to obtain a spatial pyramid.

4.2. Optimization scheme

The optimization problem in Eq. (4) can be naively formulated as the following Quadratic Programming (QP) problem:

$$\begin{aligned} \min_{\boldsymbol{z}_i,\beta_j,z_{ij}^+,z_{ij}^-} & \|\boldsymbol{x}_i - \boldsymbol{B}\boldsymbol{z}_i\|_2^2 + \lambda \sum_{j=1}^{M} \beta_j \\ \text{s.t. } & :\beta_j \geq z_{ij}^+ - z_{ij}^- \\ & z_{ij} = z_{ij}^+ + z_{ij}^- \\ & z_{ij}^+ \geq 0, z_{ij}^- \leq 0 \\ & \text{for } i = 1, \cdots N, j = 1, \cdots, M \end{aligned} \tag{5}$$

where z_{ij}^+ and z_{ij}^- are the positive and negative part of z_{ij}, repsectively. However, it is computationally demanding to solve such a QP problem because the number of variables in our application is typically very large.

In this paper, we use a fast iterative shrinkage-thresholding (FISTA) algorithm Beck and Teboulle (2009), which is a special class of the iterative shrinkage-titleholding algorithm (ISTA) that uses an accelerated gradient like scheme. FISTA/ISTA algorithm requires the efficient solving of a subproblem in Eq. (13). In this subsection, we will first describe the FISTA/ISTA method. Then, we derive an efficient algorithm to solve our optimization problem in Eq. (10).

The ISTA algorithm is capable of solving the following problem:

$$\min_{\boldsymbol{z}} F(\boldsymbol{z}) = f(\boldsymbol{z}) + g(\boldsymbol{z}), \tag{6}$$

where $f(\cdot)$ is a smooth convex function, and $g(\cdot)$ is a continuous convex function which can be possibly nonsmooth.

The ISTA algorithm solves the optimization problem in Eq. (6) by iteratively solving an easier problem:

$$p_L(\boldsymbol{x}) = \arg\min_{\boldsymbol{z}} \left\{ g(\boldsymbol{z}) + \frac{L}{2}\|\boldsymbol{z} - (\boldsymbol{x} - \frac{1}{L}\nabla f(\boldsymbol{x}))\|_2^2 \right\}, \tag{7}$$

where L is the step size that is less than the bound of the Hessian matrix of $f(\cdot)$. ISTA algorithm is efficient only if this subproblem can be efficiently solved.

In our opimization problem, the functions $f(.)$ and $g(.)$ in Eq. (6) are

$$f(\boldsymbol{Z}) = \sum_{i=1}^{N} \|\boldsymbol{x}_i - \boldsymbol{B}\boldsymbol{z}_i\|_2^2, \tag{8}$$

$$g(\boldsymbol{Z}) = \lambda\|\boldsymbol{\beta}\|_1, \tag{9}$$

respectively, where $\boldsymbol{\beta}$ is defined in equation (2).

Thus the first order derivative of f with respect to $\boldsymbol{z}_i$ is

$$\nabla f(\boldsymbol{Z})_{\boldsymbol{z}_i} = -2\boldsymbol{B}^T(\boldsymbol{x}_i - \boldsymbol{B}z_i)$$

and the problem $p_L(\cdot)$ in Eq. (13) can be written as

$$p_L(\boldsymbol{x}) = \arg\min_{\boldsymbol{z}} \left\{ \lambda \|\boldsymbol{\beta}\|_1 + \frac{L}{2} \|\boldsymbol{z} - (\boldsymbol{x} - \frac{1}{L} \nabla f(\boldsymbol{x}))\|_2^2 \right\}, \tag{10}$$

where each $\boldsymbol{x}, \boldsymbol{z}$ are the concatenation of the n individual vectors.

We show next that the problem $p_L(\cdot)$ can be efficiently solved, thus we can efficiently apply FITSA algorithm to solve our optimization problem. The rest of this subsection is devoted to the derivation the solution to the problem $p_L(.)$.

In order to solve the problem (10), we notice that the two parts of the the objective function

$$\|\boldsymbol{\beta}\| = \sum_{i=1}^{M} \max\{|z_{i1}|, |z_{i2}|, \cdots, |z_{iN}|\} \tag{11}$$

and

$$\frac{L}{2} \|\boldsymbol{z} - (\boldsymbol{x} - \frac{1}{L} \nabla f(\boldsymbol{x}))\|_2^2 \tag{12}$$

can be decomposed as the linear combination of M functions of of $z_{i1}, z_{i2}, \cdots, z_{i2}$, with $i = 1, 2, \cdots,$.

After decomposition, we can solve this problem by solving the subproblems separately in the following form

$$\min_{\boldsymbol{z}} \frac{1}{2} \|\boldsymbol{z} - \boldsymbol{y}\|_2^2 + \lambda \|\boldsymbol{z}\|_\infty, \tag{13}$$

where $\boldsymbol{z}$ is the variable of each subproblem, i.e., $z_{i1}, z_{i2}, \cdots, z_{i2}$, and $\boldsymbol{y}$ is the corresponding vector of $\boldsymbol{x} - \frac{1}{L} \nabla f(\boldsymbol{x})$ in Eq. (12).

Because the norm function is convex, Eq. (13) is a convex optimization problem. However, the nonsmoothness of the problem makes it nontrivial.

Suppose that at the optimal point, we have N elements in $\boldsymbol{z}$, denoted by $z_{s_1}, \cdots, z_{s_N}$ whose absolute values are the maximum absolute value of $\boldsymbol{z}$, denoted z^*. The objective function can be transformed to

$$\frac{N}{2}(z^* - \frac{1}{N} \sum_{i=1}^{N} |y_{s_i}|)^2 + \lambda z^* + \frac{1}{2} \sum_{j \in \text{others}} (z_j - y_j)^2 + \text{const} \tag{14}$$

Due to the properties of a quadratic function, the optimal point should be

$$\begin{aligned} z_{s_i} &= \text{sign}(y_{s_i}) \frac{1}{N} (\sum_{i=1} |y_{s_i}| - \lambda)^+, \quad i = 1, \cdots, N \\ z_j &= y_j, \qquad\qquad\qquad\qquad\qquad\qquad j \in \text{others} \end{aligned} \tag{15}$$

where

$$(y)^+ = \begin{cases} y & \text{if } y > 0 \\ 0 & \text{if } y \leq 0 \end{cases}$$

By definition, we should have all $|y_j| \leq z^*$; otherwise, z^* would not be the maximum absolute value. In addition, $|y_{s_i}| \geq z^*$; otherwise, we have a better solution by choosing $|z_{s_i}| = |y_{s_i}|$.

Without the loss of generality, we assume $|y_1| \geq |y_2| \geq \cdots \geq |y_M|$ and $z^* > 0$. We need to find N such that

$$|y_{N+1}| \leq \frac{1}{N}(\sum_{i=1} |y_i| - \lambda) \leq |y_N|.$$

This is equivalent to

$$\sum_{i=1}^{N}(|y_i| - |y_N|) \leq \lambda,$$

and

$$\sum_{i=1}^{N+1}(|y_i| - |y_{N+1}|) \geq \lambda.$$

Since $\Phi(N) = \sum_{i=1}^{N}(|y_i| - |y_N|)$ is a non-decreasing function of N, and $\Phi(1) = 0 \leq \lambda$, if $\Phi(M) \geq \lambda$, there must exist $1 \leq N < M$ that satisfies these conditions. Otherwise, we can choose $N = M$.

In addition, because of the non-decreasing property of the function $\Phi(N)$, we can find N by a bipartite search, and find the optimal value of our sub-problem efficiently.

5. Dictionary Learning

Although using the dictionary generated by off-the-shelf clustering methods such as the k-means algorithm can usually achieve acceptable performance, we should expect training a dictionary that fits better in our sparse coding scheme to improve the performance. Moreover, most clustering algorithms require the loading of all the data into the memory, which is not suitable for very large datasets. In this section, we design an online learning algorithm similar in spirit to the one proposed in Mairal et al. (2010) for dictionary learning. This approach can learn a dictionary that achieves better performance and can be applied to very large datasets.

By enforcing the sparsity of the pooled vectors, we encourage the basis in the dictionary to be more discriminative. Therefore, optimizing the proposed objective function (16) can obtain a better dictionary, which will be demonstrated in our experiments in section 6.

The proposed dictionary learning method considers the following optimization problem that extends Eq. (4):

$$\begin{aligned} &\min_{\boldsymbol{B},\boldsymbol{z}_i^j} \sum_{j=1}^{I}\sum_{i=1}^{N} \|\boldsymbol{x}_i^j - \boldsymbol{B}\boldsymbol{z}_i^j\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \\ &s.t. \|\boldsymbol{b}_j\|^2 \leq 1, \end{aligned} \tag{16}$$

where $\boldsymbol{b}_j$ is a column of $\boldsymbol{B}$ and is the basis in the dictionary; $\boldsymbol{x}_i^j$ and $\boldsymbol{z}_i^j$ are the i-th training data and features of the j-th subregion, respectively. Eq. (16) can be solved by coordinate descent methods that iteratively optimize the objective function with respect to the variable $\boldsymbol{D}$ and $\boldsymbol{z}_i^j$. The details of the algorithm are given in Algorithm 1.

An very important issue in dictionary learning is that some basis may be rarely used in sparse coding, resulting in a singular $\boldsymbol{C}$ in Algorithm 1. In this case, we remove these basis from the

dictionary, and replace them by a random data samples that are sufficiently far away from these basis. This step is very important for successful dictionary learning.

Algorithm 1: Online Dictionary Learning
Take λ (regularization parameter), $\boldsymbol{B}_0$ (initial dictionary), T (number of iterations).
Set $\boldsymbol{C}_0 = 0, \boldsymbol{D}_0 = 0$
for $t = 1$ *to* T **do**
  Draw the training data $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N$ from one region of a single image.
  Compute $\boldsymbol{z}_1, \cdots, \boldsymbol{z}_N$ by solving problem (4).
  Set $\boldsymbol{C}_t = \boldsymbol{C}_{t-1} + \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{z}_i \boldsymbol{z}_i^T$.
  $\boldsymbol{D}_t = \boldsymbol{D}_{t-1} + \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{z}_i^T$
  Denote $\boldsymbol{B}_{t-1} = [\boldsymbol{b}_1, \cdots, \boldsymbol{b}_M], \boldsymbol{C} = [\boldsymbol{c}_1, \cdots, \boldsymbol{c}_M], D = [\boldsymbol{d}_1, \cdots, \boldsymbol{d}_M]$
  repeat
    for $j = 1$ *to* M **do**
      Update the j-th column of $\boldsymbol{B}$ to optimize (16).

$$\begin{aligned} \boldsymbol{u}_j &= \frac{1}{C[j,j]}(\boldsymbol{d}_j - \boldsymbol{B}\boldsymbol{c}_j) + \boldsymbol{b}_j, \\ \boldsymbol{b}_j &= \frac{1}{\max(\|\boldsymbol{u}_j\|_2, 1)}\boldsymbol{u}_j \end{aligned} \tag{17}$$

    end
  until *convergence*;
  Set $\boldsymbol{B}_t$ to the obtained matrix
end
return $\boldsymbol{B}_T$

6. Experiments

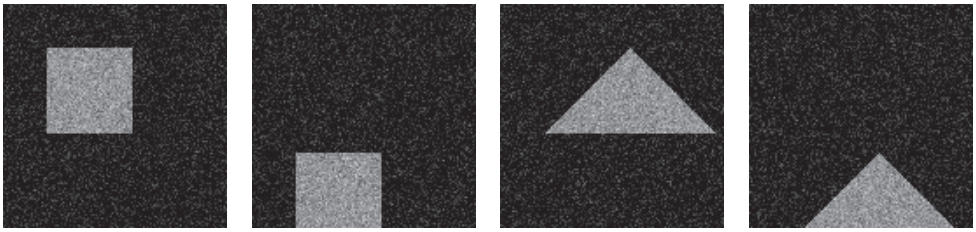
In this section, we evaluate and validate the proposed sparse coding and dictionary learning approach on synthetic data and Caltech-101 dataset Feifei et al. (2007). The experimental results demonstrate that the proposed sparse coding scheme is able to learn a better dictionary and achieve a better performance for image classification.

6.1. Synthetic data

To illustrate the idea and validate the performance of our spatial locality-aware sparse coding and dictionary learning scheme, we first evaluate our scheme on synthetic data.

Two types of synthetic images are generated. All of the generated images have a black background and are of size 128×128 pixels. Each first type image has a rectangle in it, and a second-type image has a triangle in it. We shift the rectangles and triangles to generate 9 images for each type. The images are corrupted by zero-mean white Gaussian noise with variance 0.04. Two examples for each type of image are shown in Fig. 4. This dataset is used to illustrate and validate our sparse coding and dictionary learning scheme.

The raw-patch features are extracted from each images on a 16×16 pixels sliding window. All the vectors are normalized to norm-1. First, the K-means algorithm is used to cluster all the



(*a*) (*b*) (*c*) (*d*)

Figure 4: The example of the synthetic data: the left images are rectangle images, the right images are the triangle images

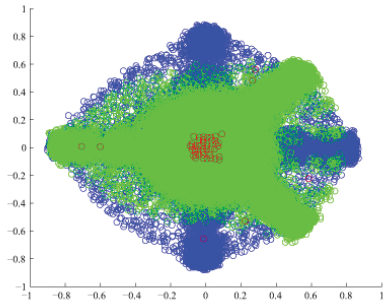


Figure 5: The embedding of the raw-patch features of the synthetic images and the cluster centers generated by the k-means algorithm, where the blue points are the feature points of rectangle images; the green point are the feature points of the triangle images; and the red points are cluster centers.

raw-patch features to 50 classes. We denote the basis of the corresponding dictionary as the centers of each class. The embedding result of the basis is illustrated in Fig. 5, where the blue points are the feature points of the rectangle images; the green point are the feature points of the triangle images; and the red points are basis. It can be noticed that most of the bases lie in the places where the raw-patch features correspond to the black background, because the majority of the extracted patches are from the background. An illustration of the cluster centers is shown in Fig. 6. Most of bases are noises in the background, which are not useful for classification.

The dictionary learned by the traditional sparse coding dictionary learning is show in Fig. 7. This dictionary is not very different from the dictionary generated by the k-means algorithm.

Then, we apply our dictionary learning algorithm to learn a new dictionary. The embedding illustration of the learned dictionary is shown in Fig. 8. Compared to the dictionary generated by the k-means algorithm, our algorithm is able to learn a dictionary whose bases are less concentrated. The illustration of the bases in the learned dictionary is shown in Fig. 9. More bases in this dictionary are on the edges of the triangles or rectangles, which are useful for classification.

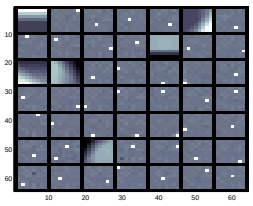


Figure 6: The illustration of all the cluster centers of the k-means algorithm on the synthetic data.

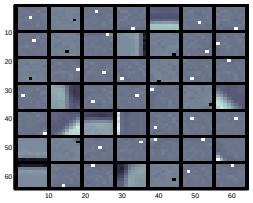


Figure 7: The illustration of all the basis of the dictionary learned with traditional sparse coding dictionary learning on the synthetic data.

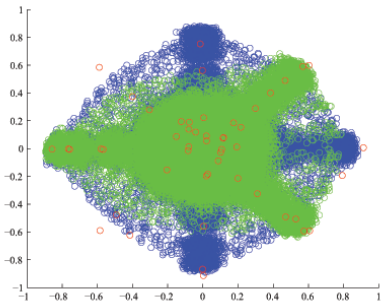


Figure 8: The embedding of the raw-patch features of the synthetic images and the basis in the learned dictionary, where the blue points are the feature points of rectangle images; the green point are the feature points of the triangle images; and the red points are cluster centers.

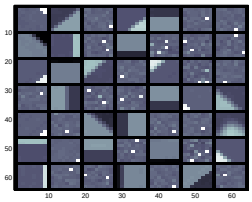
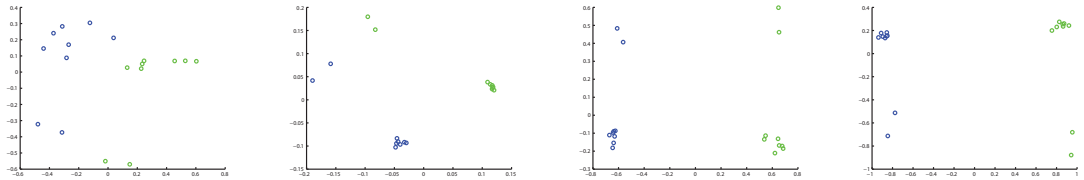


Figure 9: The illustration of all the bases in the learned dictionary on the synthetic data.



(*a*) (*b*) (*c*) (*d*)

Figure 10: The embedding of the feature vectors (from left to right): traditional sparse coding using k-means dictionary, traditional sparse coding using the learned dictionary, the proposed sparse coding using k-means dictionary, illustrates the feature vectors of the proposed sparse coding using the learned dictionary.

Finally, we evaluate the classification performance of our sparse coding scheme and the learned dictionary. Four test settings are listed below. In each case, we do not use a spatial pyramid but perform spatial pooling on the whole images.

- Conventional sparse coding using the k-means dictionary.
- Spatial locality-aware sparse coding using the k-means dictionary.
- Sparse coding using the learned dictionary.
- Spatial locality-aware sparse coding using the learned dictionary.

The classification results are shown in Table 1, in which the accuracy is represented by the average accuracy of 200 rounds classification on random permutation of the training/test images. Because our synthetic dataset is pretty easy, each case achieves 100% accuracy when more than 4 images in each class are used as training data. However, if less than 4 images are used as training data, using the learned dictionary can significantly increase the classification accuracy. In addition, our spatial locality-aware sparse coding performs better than conventional sparse coding in classification.

The embedding illustrations of all images' feature vectors from the synthetic data are shown in Fig. 10. It can be observed that the proposed sparse coding and dictionary learning algorithm are able to extract features that are more discriminative.

Number of training image	1	2	3	4
Sparse coding using the k-means dictionary	53.68	90.64	99.6	100
Spatial locality-aware sparse coding using the k-means dictionary	52.93	99.78	100	100
Sparse coding using the learned dictionary	50.56	100	100	100
Spatial locality-aware sparse coding using the learnt dictionary	100	100	100	100

Table 1: The classification accuracy comparison on the synthetic data

Number of training image	5	10	15	20	25	30
NBNN Boiman et al. (2008)	-	-	65.00	-	-	70.40
SPM Lazebnik et al. (2006)	-	-	56.40	-	-	64.60
ScSPM Yang et al. (2009)	-	-	67.0	-	-	73.2
LLC-Coding Wang et al. (2010)	**51.15**	59.77	65.43	67.74	70.16	**73.44**
Propposed algorithm using k-means dictionary	47.7	57.9	63.5	66.8	69.4	71.32
Proposed algorithm using learned dictionary	51.1	**61.8**	**67.22**	**69.4**	**71.45**	72.60

Table 2: The classification accuracy comparison on the Caltech-101 data.

6.2. Caltech-101

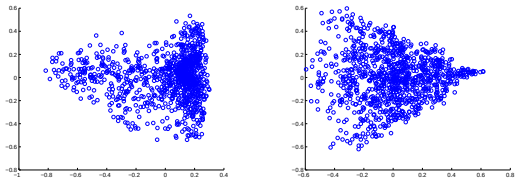
The Caltech-101 dataset contains 9144 images in 101 classes with significant intra-class variances. We use similar settings as in Wang et al. (2010) in evaluating our algorithm. In this experiment, the images were resized to be no larger than 300×300 pixels with preserved aspect ratio. As suggested by the original dataset Feifei et al. (2007), we randomly partitioned the whole dataset into $5, 10, \cdots, 30$ training images per class and no more than 50 testing images per class, and measured the performance using average accuracy over 101 classes and a "background" class.

For local feature extraction, SIFT features are extracted on 8×8 pixels size dense grid with patch size 16×16, and three levels of spatial pyramids are used. The joint-sparse coding and spatial pooling is performed in each smallest subregion.

We use the k-means algorithm and the proposed spatial locality-aware sparse coding to obtain two different dictionaries with 1024 basis. The embedding illustration of the dictionary generated by the k-means algorithm and the learned dictionary are shown in Fig. 11(*a*) and Fig. 11(*b*). Similar to the experiments in synthetic data, our algorithm is able learn a dictionary whose bases are less concentrated.

In our evaluation, we compare the image classification accuracy of the proposed sparse coding scheme using both dictionaries. It can be observed that using the learned dictionary results in more accurate classification. We also compare to other features encoding schemes such as the sparse coding Yang et al. (2009) and LLC-Coding Wang et al. (2010). The experimental results demonstrate that the proposed approach performs better when a small training dataset is used. The detailed classification results are shown in Talbe 2. The accuracy measure in this table is the average accuracy of 20 rounds classification on random permutation of the training/test images.

The proposed algorithm takes 10 seconds to encode the 1824 descriptors from an image of size 241×300, while the naive QP formulation takes more than 5 minutes to encode these descriptors. The experiments are done in a 8-core 2.27GHZ Xeon L5520 machine with 50GB memory.



(*a*) (*b*)

Figure 11: The embedding illustration of k-means clusters (a) and learnt dictionary (b) in Caltech 101 dataset.

7. Conclusion

This paper presents a novel sparse coding approach that can take spatial locality information into account and proposes an efficient optimization algorithm to solve it. The proposed spatial locality-aware sparse coding is able to obtain more stable sparse vectors and learn a better dictionary by regularizing the sparseness of the max-pooled vectors in sub-regions. Experiments on synthetic and benchmark dataset demonstrate the effectiveness of the proposed framework. Future directions of the work may include the following issues. First, given the usefulness of the spatial information, how can we exploit richer spatial information? Second, supervised and semi-supervised approaches can be explored to improve the dictionaries.

References

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009. URL http://iew3.technion.ac.il/~becka/papers/71654.pdf.

Oren Boiman, Eli Shechtman, and Michal Irani. In defense of Nearest-Neighbor based image classification. In *CVPR*, number i. Ieee, June 2008. ISBN 978-1-4244-2242-5. doi: 10.1109/CVPR.2008.4587598. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4587598.

L Feifei, R Fergus, and P Perona. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1):59–70, April 2007. ISSN 10773142. doi: 10.1016/j.cviu.2005.09.012. URL http://linkinghub.elsevier.com/retrieve/pii/S1077314206001688.

R Fergus, P Perona, and A Zisserman1. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, pages 264–271, 2003. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1211479.

Shenghua Gao, I. Tsang, and L.T. Chia. Kernel Sparse Representation for Image Classification and Face Recognition. *ECCV*, (i):1–14, 2010a. URL http://www.springerlink.com/index/68671G96035W3777.pdf.

Shenghua Gao, I.W.H. Tsang, L.T. Chia, and Peilin Zhao. Local features are not lonelyLaplacian sparse coding for image classification. In *CVPR*, pages 3555–3561. IEEE, 2010b. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5539943.

K Grauman and T Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *ICCV*. IEEE Computer Society, 2005. URL http://www.computer.org/portal/web/csdl/doi/10.1109/ICCV.2005.239.

S. Lazebnik, C. Schmid, and J Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, volume 2. IEEE, 2006. ISBN 0769525970. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1641019.

Honglak Lee, A. Battle, R. Raina, and A.Y. Ng. Efficient sparse coding algorithms. In *NIPS*, volume 19, page 801. Citeseer, 2007. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.69.2112&rep=rep1&type=pdf.

Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. *ICML*, 2009. doi: 10.1145/1553374.1553463. URL http://portal.acm.org/citation.cfm?doid=1553374.1553463.

Julien Mairal, Francis Bach, Inria Willow Project-team, and Guillermo Sapiro. Online Learning for Matrix Factorization and Sparse Coding. *JMLR*, 11:19–60, 2010.

Ignacio Ramirez, Pablo Sprechmann, and Guillermo Sapiro. Classification and clustering via dictionary learning with structured incoherence and shared features. In *CVPR*, number 1. IEEE, 2010. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5539964.

Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong. Locality-constrained Linear Coding for Image Classification. In *CVPR*, 2010.

Jianchao Yang, Kai Yu, Yihong Gong, and Thomas Huang. Linear Spatial Pyramid Matching Using Sparse Coding for Image Classification. In *CVPR*, 2009.

Jianchao Yang, Kai Yu, and Thomas Huang. Supervised Translation-Invariant Sparse Coding. In *CVPR*, 2010a.

Jianchao Yang, Kai Yu, and Thomas Huang. Efficient Highly Over-Complete Sparse Coding using a Mixture Model. *ECCV*, pages 113–126, 2010b. URL http://www.springerlink.com/index/KH22371487501W47.pdf.

Kai Yu and T. Zhang. Improved Local Coordinate Coding using Local Tangents. In *ICML*. Citeseer, 2010. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.169.3105&rep=rep1&type=pdf.

Kai Yu, T. Zhang, and Y. Gong. Nonlinear learning using local coordinate coding. In *NIPS*, volume 22. Citeseer, 2009. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.157.9788&rep=rep1&type=pdf.

Junsong Yuan and Ying Wu. Context-Aware Clustering. In *CVPR*, 2008.

Xi Zhou, Kai Yu, Tong Zhang, and Thomas S Huang. Image Classification using Super-Vector Coding of Local Image Descriptors. In *ECCV*, 2010.

Wangmeng Zuo, Zhouchen Lin, Zhenhua Guo, and David Zhang. The multiscale competitive code via sparse representation for palmprint verification. In *CVPR*, pages 2265–2272. IEEE, 2010. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5539909.