Sparse Spatial Coding: A Novel Approach for Efficient and Accurate Object Recognition

Gabriel L. Oliveira Erickson R. Nascimento Antonio W. Vieira Mario F. M. Campos

Abstract— **Successful state-of-the-art object recognition techniques from images have been based on powerful methods, such as sparse representation, in order to replace the also popular vector quantization (VQ) approach. Recently, sparse coding, which is characterized by representing a signal in a sparse space, has raised the bar on several object recognition benchmarks. However, one serious drawback of sparse space based methods is that similar local features can be quantized into different visual words. We present in this paper a new method, called Sparse Spatial Coding (SSC), which combines a sparse coding dictionary learning, a spatial constraint coding stage and an online classification method to improve object recognition. An efficient new off-line classification algorithm is also presented. We overcome the problem of techniques which make use of sparse representation alone by generating the final representation with SSC and max pooling, presented for an online learning classifier. Experimental results obtained on the Caltech 101, Caltech 256, Corel 5000 and Corel 10000 databases, show that, to the best of our knowledge, our approach supersedes in accuracy the best published results to date on the same databases. As an extension, we also show high performance results on the MIT-67 indoor scene recognition dataset.**

I. INTRODUCTION

Recognizing objects in images has been a challenging task and, for a good number of years, it has attracted the attention of a large number of researchers from several communities such as robotics, computer vision and machine learning. Object recognition is at the core of important tasks, like tracking and Simultaneous Localization And Mapping (SLAM). Unquestionably, the amount of information captured in a single frame has naturally been conducive for the numerous image based techniques and has spawned several categorization methodologies.

Sparse Coding (SC) has been successfully used for image denoising [10] and image restoration [20], [22]. However, only recently SC has been effectively applied in lieu of Vector Quantization (VQ) techniques in object recognition tasks, and is now considered as the state-of-the-art [14], [35]. This powerful representation has also been regarded as likely to be separable in high-dimensional sparse spaces [27] and therefore suitable for classification. On the downside, a weak point of sparse coding is that the regularization process can select different basis for two similar patches. The observation of this fact led us to investigate ways to overcome this severe drawback of sparse representation.

The authors are affiliated with the Computer Vision and Robotic Laboratory (VeRLab), Computer Science Department, Universidade Federal de Minas Gerais, MG, Brazil. Antonio Vieira is also affiliated with CCET, UNIMONTES, MG, Brazil. This work has been supported by grants from CNPq, CAPES and FAPEMIG. E-mails: {gabriel,erickson,awilson,mario}@dcc.ufmg.br

The main problem we deal with in this paper is the uncovering of the semantic category of an image. Much of the work for whole image categorization has already been successfully accomplished using Bag-of-Features (BoF) based approaches. However, those approaches represent an image as an orderless collection of local features, that obviously do not capture global features, such as shape, to differentiate between objects. In order to overcome this loss of spatial information, an extension to BoF called Spatial Pyramid Matching (SPM), was proposed by Lazebnik et al. [16]. Nowadays, SPM is an important component of state-of-the-art object recognition techniques such as [6], [9], [11], [32], [35].

Indeed, SPM is a preferred choice, since it creates geometrical relationships between features, which combined with SC, leads to high accuracy results. Therefore, our work aims to develop a new approach for object recognition called SSC, which combines the advantages of SPM and implements a spatial Euclidian coding representation to overcome the aforementioned SC drawbacks.

Our method is composed of three main steps: i) a Training Phase, ii) a Coding Phase, and iii) an Online Learning approach to predict the labels that will be assigned to a feature.

In the training phase we build the dictionary with a set of random patches extracted from the training image set. These patches are normalized and passed on to the dictionary learning process.

The Coding Phase can be further divided into two steps: i) the extraction of local descriptors, which can use SIFT [18] or SURF [2] descriptors, and ii) code generation, based on the dictionary and on the quantization of each descriptor, using a spatial constraint, instead of sparsity. The codes associated with each region are pooled together to form a global image signature. We maximize the signature using max pooling [26].

The final stage of our method sends the global features to an Online Classification method. We chose online learning based on the requirements that are typical of real robots: i) small memory availability; ii) large amounts of data, and iii) suitability for data streaming (typical of several robotic tasks).

Online learning is well suited to several robotic tasks where in general the robot does not have access to the entire data domain. This is also similar to decision making problems, where parts of the data are incrementally presented

over time [29]. This idea can be exemplified by a simple game quiz. Consider a student and a teacher executing together, n times, the following steps:

1) An input sample is presented to the student.
2) The student responds to the input with a prediction.
3) The teacher reveals the true answer for the input.
4) If the prediction is correct, then the model is reinforced, if it is wrong, the student is penalized and updates his model based on the correct information.

The goal of the student is to minimize the cumulative error over time by updating his internal model of the problem.

We propose here a new method to improve object recognition called (SSC), which combines the learning of a sparse coding dictionary, a spatial constraint coding stage and an online classification method. Furthermore, we also propose a new and efficient off-line algorithm. Experimental results, presented later in this paper, show that, to the best of our knowledge, the results obtained on the Caltech 101, Caltech 256, Corel 5000 and Corel 10000 datasets, achieve accuracies that are superior to the best published results to date on the same databases. In addition, we also obtain high performance results on the MIT-67 indoor scene recognition dataset.

The main contributions of this paper are: i) A new object recognition method which makes use of SC for dictionary learning and a coding stage based on spatial constraint, ii) a new off-line method based on SVD, called Orthogonal Class Learning OCL, designed to take advantage of the high dimensionality of features when compared to the number of feature examples, and iii) an object recognition technique based on online learning, that when combined with the previous steps, leads to state-of-the-art performance results on several popular benchmark datasets.

After discussing the related works in the next section, in Section III we present our methodology to learn dictionaries and to generate code, based on sparse coding and locality, respectively. We also present a novel off-line classification method and an online learning algorithm that comprise the final classification technique of our methodology. Experimental results are described in Section IV followed by Section V, which reports on what we have concluded with this investigation, the work underway and the possible next research directions.

II. RELATED WORK

Several approaches using SC and dictionary learning for image classification have been proposed in recent years. Some of these approaches use supervised feature learning [1], [6], [14], [37], [38] which achieve high performance results on several object/scene datasets. Unsupervised learning [11], [16], [32], [35] was used to recognize images on a large-scale, unlabeled database.

Two recent works dealing with SC and supervised dictionary learning are [14] and [6]. [14] proposes a supervised dictionary learning technique called Label Consistent KSVD (LC-KSVD), that assigns labels (a column of the dictionary matrix), in order to increase the discrimination power in sparse coding during the learning process of a dictionary. This method combines dictionary learning and a single predictive linear classifier with an objective learning function. In [6], the authors proposed a method for supervised dictionary learning with a deep analysis of coding and spatial pooling modules. This evaluation ushered two discoveries: First, that sparse coding improves soft quantization, and second, that max pooling, almost in all the studied cases, is superior to average pooling, which is unequivocally perceived when using a linear SVM.

Another research stream is related to unsupervised dictionary learning for object recognition. Recent works have proposed additional regularization and/or constraints such as spatial properties, like [11], [15] and [32]. Yang et al. [35] propose an extension to the SPM method [16] by replacing the vector quantization with a sparse coding approach. After running SPM, a max pooling technique is applied to summarize all the local features representing the image. By incorporating locality, the approach in [32] aims to decrease the reconstruction error of the sparse coding based on the idea that similar patches will have similar codes given the locality.

Our approach could be classified as an unsupervised dictionary learning technique, and more specifically, it shares some similarities with the work of [11], [32]. However, instead of using locality for training the dictionary and the to generate coding, our method uses sparse representation for the dictionary, since our data for training is limited. As Coates et al. [8] conclude, sparse coding achieves consistent results when a small number of examples are available. The work of [28], presents a thorough analysis of the importance of sparse representation for image classification, also points out the relevance of sparsity for learning the feature dictionary. We then move on to an encoding process which uses spatial similarity, initially in an off-line mode, and finally in an online classification approach. The details of our method will be described in the following sections.

III. METHODOLOGY

As we have sketched before, our method is composed by a training phase, corresponding to sparse coding dictionary learning, and a coding phase, corresponding to the sparse spatial coding process. After coding, our SSC image signature is presented to a learning method that could be our offline OCL or our online LaRank. In what follows we will discuss each of the main modules of the SSC technique.

A. Dictionary Learning

First, we compute a set X of SIFT descriptors $x_i, \forall i = 1, \ldots, N$ from a random collection of image patches and use their information to solve

$$\underset{U,D}{\operatorname{argmin}} \sum_{m=1}^{N} \|x_m - \mu_m D\|^2 + \lambda|\mu_m|, \qquad (1)$$

where $U = [\mu_1 \ldots \mu_m]$ is the set of basis of each descriptor.

Equation 1 presents the problem of not being convex simultaneously for both U and V. Honglak et al. [17] also

points out this problem and propose a technique, consisting of optimizing one variable while the other remains constant. For example, when U is constant, D becomes convex and vice-versa.

When the dictionary D is fixed, Eq. 2 can be rewritten as:

$$\underset{\mu m}{\operatorname{argmin}} \|x_m - \mu_m D\|_2^2 + \lambda|\mu_m|, \tag{2}$$

where the $\|x_m - \mu_m D\|_2^2$ constraint describes the reconstruction error, and λ is a regularization parameter used to prevent overfitting.

As stated, this problem is known as Lasso, a linear regression with $L1$ norm regularization on the coefficients. It can be solved using tools such as those provided by the recently published Sparse Modeling Library (SPAMS) [21] or with a feature-sign search algorithm [17].

When U is fixed, the problem is reduced to a Least Squares problem with quadratic constraints:

$$\underset{D}{\operatorname{argmin}} \|X - UD\|_F^2 \tag{3}$$
$$\text{s.t. } \|D_k\| \leq 1, 1 \leq k \leq n,$$

which can be solved with the Lagrange Dual procedure.

Several tests were performed by extracting SIFT descriptors from random patches to train the dictionary, and iterating Eq. 2 and Eq. 3.

Finally, after the dictionary is trained, the next step is the coding phase. For that we use a spatial constraint.

B. Coding Phase

Instead of coding with a sparsity constraint, we have chosen to use the spatial Euclidean similarity, based on the works of [32] and [36]. Those works suggest that locality produces better signal reconstruction.

In VQ, each descriptor is represented by a single base. However, spatial approaches use multiple basis in order to capture possible correlations between similar descriptors.

Another factor which led us to opt for this type of coding, is also presented in the works of [32] and [36]: locality imparts a higher probability of selecting similar basis for similar patches. This is different from a SC approach, where regularization can select quite diverse basis for similar patches (see Figure 1).

Coding with spatial Euclidean similarity transforms Eq. 1 into:

$$\underset{\mu}{\operatorname{argmin}} \sum_{i=1}^{N} \|x_i - D\mu_i\|^2 + \lambda \|d_i \odot \mu_i\|^2 \tag{4}$$
$$\text{s.t. } \mu_i = 1, \forall i, i = 1, .., N,$$

where d_i is the spatial similarity member computed as

$$d_i = dist(x_i, D), \tag{5}$$

where $dist(x_i, D)$ is a vector of Euclidean distances between each input descriptor x_i and the basis of the dictionary.

Given these distances, we apply a KNN method that returns the K most similar basis for the given input, implying on a low computational requirement for our coding process. The values of d_i are normalized by the max distance to adjust for the range of possible numbers represented within the interval $(0, 1]$.

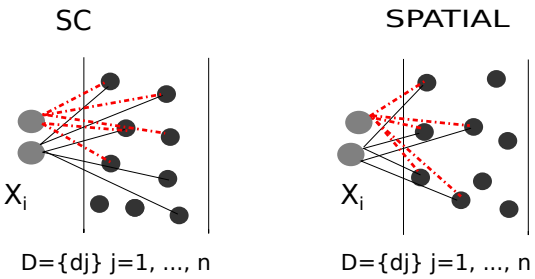

Fig. 1. One SC shortcoming is that regularization can select different basis for similar patches, a problem that spatial constraint techniques are able to overcome. X_i represent the input features and D represents the dictionary. As it can be seen in this example, the spatial Euclidean coding selected the nearest basis in the dictionary.

After coding each local feature, we perform the max pooling to concatenate each code into a final image representation. This final representation is used in one of the classification methods. We test the final signature in the OCL classification technique, and then we use an online approach that consitutes the final step of the methodology.

C. Off-line learning method

We propose an off-line method based on SVD, called Orthogonal Class Learning (OCL). This new methodology takes advantage of the high dimensionality of the feature vectors when compared to the number of feature examples, i.e., we have a set with m n-dimensional feature vectors where $n \gg m$. In this case, a base with only m components is used to represent new feature vectors. In addition, we obtain a new base for which new feature vectors are unit vectors, and pairwise orthogonal.

Consider, initially, that we have h classes, each of which represented by k n-dimensional feature vectors so that we have $m = h \times k$ feature vectors. Let $f_1, f_2, \ldots, f_m$ denote feature vectors of all training data, and let F denote the $n \times m$ matrix where columns are formed by the feature vectors, that is:

$$F = (f_1, f_2, ..., f_m). \tag{6}$$

Using SVD decomposition, we have $F = USV^T$. Instead of forming new basis from the columns of U as it is done in standard PCA, we lay hold of the fact that

$$V^T = (S^T S)^{-1} S^T U^T F, \tag{7}$$

and form a new basis $A = (S^T S)^{-1} S^T U^T$ such that in this new basis our new feature vectors are columns from V^T, which are unit vectors and pairwise orthogonal. The advantage of this new representation is that, given object classes i and j and their feature vectors as matrices F_i and

F_j formed with columns from F, we obtain new matrices $C_i = A.F_i$ and $C_j = A.F_j$, such that columns of C_i and C_j are pairwise orthogonal vectors.

Finally, we build our classifier based on the aforementioned observations. Given an object and its feature vector f, we obtain a new feature vector $e = A \times f$. The decision over the class S is given by

$$S = \underset{s}{\operatorname{argmax}} \|C_s^T e\|. \quad (8)$$

D. Online learning method

In summary, the approach used in our final methodology was based on an Online LaRank [5]. We have selected a LaRank multi-class solver. The LaRank algorithm is grounded in a randomized exploration, inspired by the perceptron algorithm [4].

LaRank was selected as solver chiefly for these following reasons:

- It reaches the equivalent accuracy levels of other SVM solvers, like SVMstruct [31], but with higher computational performance;
- It generalizes better than perceptron-based algorithms;
- It achieves nearly optimal test error rates after a single pass over the randomly reordered training set.

The Online LaRank technique achieves the same test accuracy of batch optimization after a single epoch thanks to the Reprocess step implementation over SMO-Optimization algorithm [24].

IV. EXPERIMENTS

For evaluation purposes, we tested our method in two scenarios. First, we performed experiments using our technique with off-line classification methods, such as SVM and with the OCL approach. In a second phase we tested our final methodology (Sparse dictionary learning plus spatial Euclidean coding) with an online learning algorithm.

The first test aims to show that only a sparse spatial constraint approach can lead to state-of-the-art results with our off-line approach. Furthermore, the combination of sparse coding and locality with the correct online learning method can produce superior results.

A. Parameters Setting

One of the most critical settings for an object recognition method is the choice of a local feature to be used. In our experiments, we chose SIFT [18] due to its high accuracy on several object recognition tasks [3], [16], [35]. Because of the dense grid sampling in the step for selecting regions of interest, our experiments use 6 pixels step between each region with a patch size of 16×16 pixels. During our trials, we tested the system with smaller step sizes, but as expected, the computational cost was prohibitive. We also resized the images to 300×300 pixels.

We trained all the dictionaries for the tests with 1024 basis and 20000 random patch samples. The main parameter setting for the dictionary training is the sparsity/regularization that we empirically set to $\lambda = 0.30$. For the coding stage, after experimenting with other values, we selected $K = 5$ neighbors for the KNN.

All results report the average for 10 runs with random selection of training and testing sets.

B. Off-line Methods Evaluation

We first test how the SSC method works in an off-line standard classification method, like a Linear SVM and Random Forests. Additionally, we also present the results we obtained with the new OCL algorithm. The dataset used was the Caltech 101, consisting of 101 classes with broad shape variation. For evaluation purposes, we compare our results with three recently proposed methods: LLC [32], ScSPM [35], and NBNN [3]. Table I summarizes the results. As it can be seen, our technique presents better performance, specially when combined with the OCL technique, outperforming the best results reported in the literature.

TABLE I
OFF-LINE METHODOLOGIES. WE SEE THAT THE COMBINED APPLICATION OF THE TRAINING AND CODING STEPS WITH OUR OCL APPROACH OUTPERFORMS THE RESULTS OF PREVIOUSLY REPORTED METHODS.

N. train	5	10	15	20	25	30
NBNN [3]	-	-	65.00±1.14	-	-	70.40
ScSPM[35]	-	-	67.00±0.45	-	-	73.20±0.54
LCC[32]	51.15	59.77	65.43±0.45	67.74	70.16	73.44
Ours(RF)	33.69	41.54	46.12±0.98	49.80	52.08	54.24±0.59
Ours(SVM)	46.34	56.51	61.97±0.43	65.29±0.91	67.66±0.57	71.18±0.53
Ours(OCL)	**56.70**	**65.26**	**69.00±0.74**	**71.7±0.72**	**73.62±0.51**	**75.67±0.52**

C. Online Learning Evaluation

In order to compare ours with other online learning approaches, we conducted experiments with our complete methodology and with three other online methods: ORF (Online Random Forest) [30], OMCGB (Multi-Class Gradient Boost), and OMCLPB (Online Multi-Class LPBoost) [29]. Table II reports the average results of 10 runs with randomly chosen samples from the Caltech 101 dataset for training and for testing. Each algorithm run for 10 epochs.

TABLE II
ONLINE LEARNING RESULTS. THE RESULTS SHOW A CLEAR ADVANTAGE OF OUR METHOD OVER OTHER ONLINE LEARNING TECHNIQUES BY A MARGIN EXCEEDING 21%.

training images	Ours	ORF [30]	OMCGB [29]	OMCLPB [29]
5	**55.64 ± 1.03**	43.21 ± 1.25	43.68 ± 1.52	43.95 ± 1.63
10	**65.52 ± 0.74**	47.75 ± 0.94	47.84 ± 1.54	48.4 ± 0.78
15	**69.98 ± 0.86**	49.8 ± 0.84	49.94 ± 0.75	51.09 ± 0.89
20	**73.99 ± 2.1**	41.93 ± 3.79	52.48 ± 0.81	53.25 ± 0.89
25	**75.49 ± 0.62**	53.79 ± 0.40	53.25 ± 0.49	54.87 ± 0.49
30	**77.59 ± 0.46**	55.54 ± 0.55	55.09 ± 0.85	56.57 ± 0.85

To complete the experimental tests, we analyze the behavior of each online classifier over 10 epochs. One well known problem related to online learning is that the order in which data is presented to algorithm can severely impact

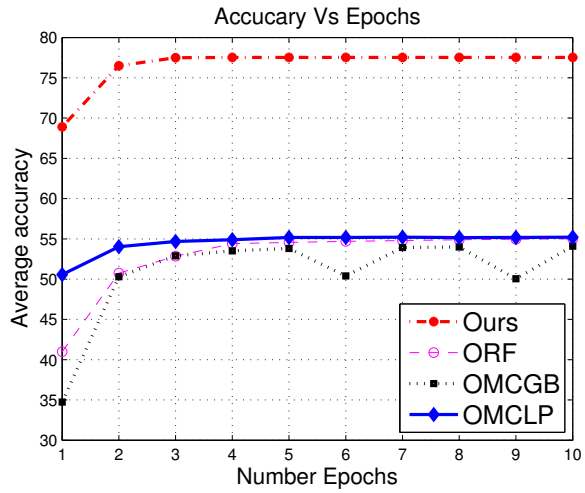

Fig. 2. Accuracies obtained under 10 epochs, average of 10 runs, on Caltech 101 with 30 images per category for training. Our method and OMCLP need 3 epochs to stabilize, but our method reaches to state-of-the-art results with a single epoch.

the performance [7]. So, to avoid such detrimental issue, we randomly shuffle our training data into n slots (10 in our trials) passing through all of them once, which constitutes one epoch. We then repeat this procedure for 10 epochs.

Figure 2 shows how each method benefits from revisiting the training set. We can see that both our method and OMCLP reach stability with only 3 epochs, however our technique presents an accuracy that is 21% better when compared to the second best algorithm. After just one epoch, the algorithm already outperforms the best results of previously published works, using a single feature.

D. Caltech 101

Caltech 101 contains 9144 images divided into 102 classes, 101 object classes and one background class, with large shape variation. The per-class number of images range from 31 to 800. To make the comparison as fair as possible, we follow the same steps of [16]. We run 10 times with different randomly selected training and test images, and the average of per-class recognition rates is recorded for each run.

Table III shows the results and the comparison with other methods recently proposed. These results confirm the hypothesis of [8], that for datasets with a small number of available training examples, for instance 30 images in 80, the sparse representation is superior to soft-thresholded ones.

Our results, shown in Table III, were based on a dictionary of 1024 basis. Nevertheless, during the experiments we have tested different dictionary sizes, such as 1024, 2048 and 4096. Our highest scores were obtained with a dictionary of size 4096. As it can be readily seen from Table III, the results represent a significant improvement for recognition rates, once our work, using a single feature approach, attained 80% of accuracy on the Caltech 101 dataset.

E. Caltech 256

Caltech 256 is an extension of the Caltech 101 dataset with 29780 images with 257 categories, including background. This dataset presents additional challenges when compared with Caltech 101, once intra-class variance and object location are larger.

Tests were performed with 15, 30, 45 and 60 images for training and the rest was used for testing. Each category contains at least 80 images. Table IV lists our results and those reported in the literature. It can be seen that as far as accuracy is concerned, our method outperforms the techniques to date.

TABLE IV

AVERAGE ACCURACY ON THE CALTECH 256 DATASET. OUR METHOD CLEARLY PRESENTS SUPERIOR ACCURACY RESULTS WHEN COMPARED WITH SEVERAL OTHER HIGH PERFORMANCE METHODS, SPECIALLY WHEN COMPARED WITH A METHOD WHICH APPLIES SPATIAL CONSTRAINT TO THE CODING PHASE (LSCSPM).

N. training	15	30	45	60
KSPM [37]	56.40	64.40 ± 0.80	-	-
ScSPM [35]	27.37 ± 0.51	34.02 ± 0.35	37.46 ± 0.55	40.14 ± 0.91
LScSPM [11]	30.00 ± 0.14	35.74 ± 0.10	38.54 ± 0.36	40.43 ± 0.38
Ours	**30.59 ± 0.35**	**37.08 ± 0.36**	**40.68 ± 0.16**	**43.48 ± 0.38**

F. Corel Datasets

Corel 1000, 5000 and 10000 datasets were originally created for Content-Based Image Retrieval. However, we believe that they are of particular interest to our tests, since they have a large number of images and they are based on natural images including those from outdoor scenes. The same procedure used with the Caltech 101 experiments was applied to these tests. We chose to perform experiments using 50 images for training and 50 for testing.

Table V summarizes the results on Corel datasets, with our approach compared against SMK, LCC, ScSPM and LScSPM. We highlight the greater recognition rate of our technique. Furthermore, this table demonstrates the superiority of our approach on the Corel 5000 and 10000, even when compared with a method which applies spatial constraints to the coding phase as [32]. One can observe that our method attains results comparable to state-of-the-art on the Corel 1000 dataset.

TABLE V

RESULTS IN COREL DATASETS.

Methods	Corel 1000	Corel 5000	Corel 10000
SMK [19]	77.90	-	-
LCC [32]	-	76.48 ± 0.77	67.72 ± 0.51
ScSPM [35]	86.20 ± 1.01	77.18 ± 0.57	68.39 ± 0.30
LScSPM [11]	**88.40 ± 0.78**	-	-
Ours	**88.40 ± 0.79**	**78.19 ± 0.63**	**69.33 ± 0.44**

G. MIT 67 Indoor

We also compare our method with the challenging scene dataset MIT 67. This dataset constitutes the largest publicly

TABLE III

RECOGNITION RESULTS ON CALTECH 101. THE RESULTS CAN BE DIRECTLY COMPARED WITH THE LITERATURE, SINCE ALL THE WORKS USE THE SAME METHODOLOGY TO PERFORM THE EXPERIMENTS. OUR METHOD (IN BOLDFACE) HAS SUPERIOR RECOGNITION RATES WHEN COMPARED WITH ALL THE SINGLE FEATURE APPROACHES FOUND IN THE LITERATURE. FURTHERMORE, WE REPORT RESULTS WITH A DICTIONARY OF 4096 BASIS, SHOWN AT OURS^{4096} LINE. FOR ALL THE CASES, OUR WORK LARGELY OUTPERFORMS THE BEST AMONG THE CURRENT SINGLE FEATURE PUBLISHED TECHNIQUES.

Number of training samples	5	10	15	20	25	30
Malik [37]	46.6	55.8	59.1	62.0	-	66.2
KSPM [16]	-	-	56.40	-	-	64.40 ± 0.80
NBNN [3]	-	-	65.00 ±1.14	-	-	70.40
ML+CORR [12]	-	-	61.00	-	-	64.14 ± 1.18
Boureau [6]	-	-	-	-	-	75.7 ± 1.1
Coates [8]	-	-	-	-	-	72.6 ± 0.9
SRC [33]	48.8	60.1	64.9	67.7	69.2	70.7
K-SVD [1]	49.8	59.8	65.2	68.7	71.0	73.2
D-KSVD [38]	49.6	59.5	65.1	68.6	71.1	73.0
ScSPM [35]	-	-	67.00	-	-	73.20
LCC [32]	51.15	59.77	65.43	67.74	70.16	73.44
LC-KSVD [14]	49.6	63.1	67.7	70.5	72.3	73.6
Ours	**55.64 ± 1.03**	**65.52 ± 0.74**	**69.98 ± 0.86**	**73.99 ± 2.1**	**75.49 ± 0.62**	**77.59 ± 0.46**
Ours^{4096}	**59.19 ± 1.14**	**68.65 ± 0.65**	**73.09 ± 0.77**	**76.18 ± 0.58**	**78.22 ± 0.43**	**80.02 ± 0.36**

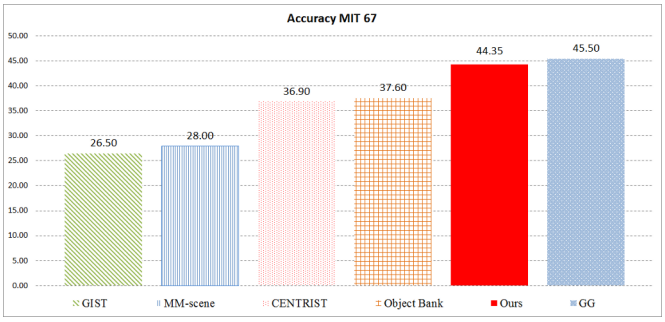

Fig. 3. Average classification rates for MIT 67 indoor scene dataset, with the exception of our result and [23], all other methods present the accuracy of a single run. Our method reaches high performance results, although inferior to GG, which shows an accuracy rate of 45.5% with a standard deviation of 1.1, against our method which presents accuracy of 44.35% with a standard deviation of 0.90.

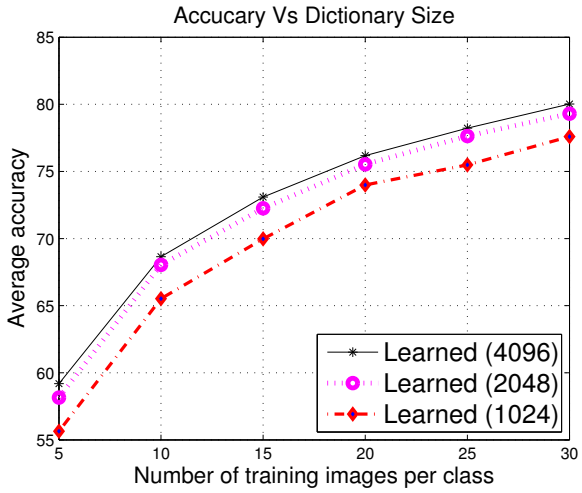

Fig. 4. Performance of different sizes of dictionaries (Caltech 101). It can be seen that there is an improvement with the increase in the size of the dictionary, but this increase topples down for 4096 basis and larger.

available benchmark base for scene recognition, with 67 classes and 15620 images. It presents large in-class variability and few distinctive attributes when compared to Scene-15 [16]. The accuracy metric is the same used in other experiments. We follow the same experimental setup of [25], which uses 80 images per class for training and 20 images per class for testing.

Figure 3 compares our results with other works reported in literature, such as GIST [25], MM-scene [39], CENTRIST [34], Object Bank [13] and GG [23]. Differently from the works used for comparison, we do not apply any annotation to the images in order to show the superior results obtained by our method. One can see that our method, using a single feature for recognition, is superior to algorithms specifically tailored for this purpose: 44.35% against 36.9% obtained by [34]. However, its overall performance does not match the highest reported result in literature (45.5 – [23]).

H. Dictionary Size

An investigation on the effects of dictionary sizes, as far as accuracy is concerned, was also performed. On one hand, a small dictionary could not provide the required discriminative power; on the other hand, large dictionaries create antagonistic histograms for images of the same class, which will not match. Three sizes were tested, 1024, 2048 and 4096. As it can be seen in Figure 4, our method presents a performance enhancement with larger dictionary sizes, but this performance boost starts to decrease for dictionary sizes of 4096 basis and up. The accuracy gain from 1024 to 2048 is 2.11%, but from 2048 to 4096 is just 0.74%. These results indicates that a policy of building even bigger dictionaries has a limit, in terms of accuracy and memory efficiency.

V. CONCLUSION

This paper presented a novel methodology for object recognition, called SSC, which uses sparse coding dictionary learning combined with a spatial Euclidean coding phase. Furthermore, one encouraging result is that our image representation works with online learning algorithms, which present some desirable properties, such as low memory usage, meaning that large amounts of data can be quickly processed and be suitable for data streaming.

Experimental results show that, to the best of our knowledge, the results obtained on the Caltech 101, Caltech 256, Corel 5000, and Corel 10000 datasets, demonstrate that our approach achieves accuracy beyond the best results for single feature previously published on the same databases. We also show high performance results on the MIT 67 indoor scene recognition dataset.

Future works will include exploring sparse supervised dictionary learning methods, which could lead to better accuracy. Other types of constraints and/or additional regularization will be investigated and other datasets will be experimented.

REFERENCES

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

[2] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *ECCV*, pages 404–417, 2006.

[3] Oren Boiman. In defense of nearest-neighbor based image classification. In *CVPR*, 2008.

[4] Antoine Bordes, Léon Bottou, Patrick Gallinari, and Jason Weston. Solving multiclass support vector machines with larank. In *ICML*, pages 89–96, 2007.

[5] Antoine Bordes, Nicolas Usunier, and Léon Bottou. Sequence labelling svms trained in one pass. In *Machine Learning and Knowledge Discovery in Databases: ECML PKDD 2008*, pages 146–161, 2008.

[6] Y. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. In *CVPR*, 2010.

[7] W. Chu, M. Zinkevich, L. Li, A. Thomas, and B. Tseng. Unbiased online active learning in data streams. In *KDD*, 2011.

[8] A. Coates and Ng. Andrew. The importance of encoding versus training with sparse coding and vector quantization. In *ICML*, 2011.

[9] A. Coates, H. Lee, and Ng. Andrew. An analysis of single-layer networks in unsupervised feature learning. In *AISTATS*, 2011.

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

[11] Shenghua Gao, Ivor Wai hung Tsang, Liang tien Chia, and Peilin Zhao. Local features are not lonely? laplacian sparse coding for image classification. In *CVPR*, 2010.

[12] P. Jain, B. Kullis, and K. Grauman. Fast image search for learned metrics. In *CVPR*, 2008.

[13] Li jia Li, Hao Su, Eric P. Xing, and Li Fei-fei. Object bank: A high-level image representation for scene classification and semantic feature sparsification. In *NIPS*, 2010.

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

[15] K. Kavukcuoglu, M. Ranzato, R. Fergus, and Y. LeCun. Learning invariant features through topographic filter maps. In *CVPR*, 2009.

[16] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, pages 2169–2178, 2006.

[17] Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Y. Ng. Efficient sparse coding algorithms. In *NIPS*, pages 801–808. NIPS, 2006.

[18] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, pages 91–110, 2004.

[19] Z. Lu and H. H. Ip. Image categorization by learning with context and consistency. In *CVPR*, 2009.

[20] J. Mairal, M. Elad, and G. Sapiro. Sparse representation for color image restoration. *IEEE Transactions on Image Processing*, 17(1):53 –69, 2008.

[21] Julien Mairal. Sparse modeling software. http://www.di.ens.fr/willow/SPAMS/, August 2011.

[22] Julien Mairal, Guillermo Sapiro, and Michael Elad. Learning multiscale sparse representations for image and video restoration. Technical Report 7, 2008.

[23] Hideki Nakayama, Tatsuya Harada, and Yasuo Kuniyoshi. Global gaussian approach for scene categorization using information geometry. In *CVPR*, pages 2336–2343, 2010.

[24] John C. Platt. *Fast training of support vector machines using sequential minimal optimization*, pages 185–208. MIT Press, Cambridge, MA, USA, 1999.

[25] Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In *CVPR*, 2009.

[26] M Ranzato, Y Boureau, and Y. LeCun. Sparse feature learning for deep belief networks. In *NIPS*, 2007.

[27] Marc'Aurelio Ranzato, Christopher Poultney, Sumit Chopra, and Yann LeCun. Efficient learning of sparse representations with an energy-based model. In *NIPS*, 2006.

[28] R. Rigamonti, M. Brown, and V. Lepetit. Are sparse representation really relevant for image classification. In *CVPR*, 2011.

[29] Amir Saffari, Martin Godec, Thomas Pock, Christian Leistner, and Horst Bischof. Online multi-class lpboost. In *CVPR*, 2010.

[30] Amir Saffari, Christian Leistner, Jakob Santner, Martin Godec, and Horst Bischof. On-line random forests. In *3rd IEEE ICCV Workshop on On-line Learning for Computer Vision*, 2009.

[31] Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6:1453–1484, December 2005.

[32] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010.

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

[34] Jianxin Wu and James M. Rehg. Centrist: A visual descriptor for scene categorization. *TPAMI*, 33(8):1489 –1501, 2011.

[35] Jianchao Yang, Kai Yu, Yihong Gong, and Thomas Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, 2009.

[36] K. Yu and Y. Zhang, T. Gong. Nonlinear learning using local coordinate coding. In *NIPS*, 2009.

[37] Hao Zhang, Alexander C. Berg, Michael Maire, and Jitendra Malik. Svm-knn: Discriminative nearest neighbor classification for visual category recognition. In *CVPR*, pages 2126–2136, 2006.

[38] Qiang Zhang and Baoxin Li. Discriminative k-svd for dictionary learning in face recognition. In *CVPR*, pages 2691 –2698, june 2010.

[39] Jun Zhu, Li-Jia Li, Li Fei-Fei, and Eric P. Xing. Large margin learning of upstream scene understanding models. In *NIPS*, 2010.