Multi-class Image Classification
Sparsity does it Better

Sean Ryan Fanello[1,2], Nicoletta Noceti[2], Giorgio Metta[1] and Francesca Odone[2]
[1]*Department of Robotics, Brain and Cognitive Sciences, Istituto Italiano di Tecnologia, Via Morego 30, 16163, Genova, Italy*
[2]*DIBRIS, Università degli Studi di Genova, Via Dodecaneso 35, 16146, Genova, Italy*

Keywords: Sparse Representation, Discriminative Dictionary Learning, Object Recognition and Categorization.

Abstract: It is well assessed that sparse representations improve the overall accuracy and the systems performances of many image classification problems. This paper deals with the problem of finding sparse and discriminative representations of images in multi-class settings. We propose a new regularized functional, which is a modification of the standard dictionary learning problem, designed to learn one dictionary per class. With this new formulation, while positive examples are constrained to have sparse descriptions, we also consider a contribution from negative examples which are forced to be described in a denser and smoother way. The descriptions we obtain are meaningful for a given class and highly discriminative with respect to other classes, and at the same time they guarantee real-time performances. We also propose a new approach to the classification of single image features which is based on the dictionary response. Thanks to this formulation it is possible to directly classify local features based on their sparsity factor without losing statistical information or spatial configuration and being more robust to clutter and occlusions. We validate the proposed approach in two image classification scenarios, namely single instance object recognition and object categorization. The experiments show the effectiveness in terms of performances and speak in favor of the generality of our method.

1 INTRODUCTION

The problem of finding good representations of the data is crucial in many computer science fields. A specificity of computer vision is that input data – images or sequences of images – live in high dimensional spaces. In this case the problem of representing data "appropriately" may be formulated as a dimensionality reduction or sparse coding problem – the latter particularly important if real-time performances are required.

Over the years, there have been many attempts to design and develop compact representations of image content – mostly based on the extraction of local and interesting characteristics – to be applied to image registration, matching or object recognition. Image patches, corners, SIFT and variants (Lowe, 2004), SURF (Bay et al., 2008), HOG (Dalal and Triggs, 2005) are just a few examples.

In classification tasks, it has been shown that the sparsity of the data representations improves the overall classification accuracy – see for instance (Viola and Jones, 2004; Destrero et al., 2009) and references therein. One of the most used techniques is the so called sparse coding, first introduced in (Olshausen and Fieldt, 1997). We refer to *adaptive sparse coding* when the coding is guided from data. In this case, we require an early stage, called *dictionary learning*. The goal is to learn a basis – a set of atoms – allowing to reconstruct the input data with a small reconstruction error (Olshausen and Fieldt, 1997; Yang et al., 2009; Yang et al., 2010; Wang et al., 2010).

In this work we propose a regularized framework for data-driven dictionaries learning based on the use of a new method for sparse coding, we called *Discriminative and Adaptive Sparse Coding* (DASC). We consider a multi-class setting and build a dictionary for each class. More specifically, we propose to modify the standard dictionary learning functional by adding a new term that forces the descriptors of negative examples to be smooth and dense, as opposed to the positive examples which are constrained to have a sparser representation. The final dictionary is a collection of all the dictionaries obtained by minimizing the functionals considering each class separately. Indeed, with a sparse representation we may employ linear classifiers instead of non-linear models that usually conflict with real-time requirements. Fig. 1 gives

Fanello S., Noceti N., Metta G. and Odone F..
Multi-class Image Classification - Sparsity does it Better.
DOI: 10.5220/0004295908000807
In *Proceedings of the International Conference on Computer Vision Theory and Applications* (VISAPP-2013), pages 800-807
ISBN: 978-989-8565-47-1

an idea of the pipeline we follow in the case of application to an object classification problem.

The properties of the proposed functional not only ensure good performances with linear classifiers, but moreover can be used directly in the classification stage. Indeed, we also propose to exploit the dictionary mechanism for the classification task by classifying each single feature on the basis of the dictionary response, rather than using the reconstruction error (Yang et al., 2008; Skretting and Husy, 2006; Peyré, 2009; Mairal et al., 2008a). The main advantage of this choice is that the classification of local features allows us to deal with occlusions and presence of cluttered background.

Most of the approaches usually focus on learning dictionaries based on the *reconstruction error* (Yang et al., 2008; Skretting and Husy, 2006; Peyré, 2009), and do not exploit the prior knowledge of the classes even in supervised tasks. In (Mairal et al., 2008a) it has been proposed a discriminative method to learn dictionaries, i.e. learning one dictionary for each class. Later in (Mairal et al., 2008b) the authors extend (Mairal et al., 2008a) by learning a single shared dictionary and models for different classes mixing both generative and discriminative methods. There have been some attempts to learn invariant middle level representations (Wersing and Körner, 2003; Boureau et al., 2010), while some other works use sparse representation as main ingredient for feed forward architectures (Hasler et al., 2007; Hasler et al., 2009). Most recent works focus on learning general task purposes dictionaries (Mairal et al., 2012) or they look at the pooling stage (Jia et al., 2012) trying to learn the receptive fields that better catch all the image statistics.

In this work, we exploit the power of low-level features from a different perspective, i.e. taking advantage on the sparsity. The main contributions of our work can thus be summarized as follows

- A new functional for learning discriminative and sparse image representations, that exploits prior knowledge on the classes. Unlike other approaches, when building the dictionary of a given class, we also consider the contributes of negative examples. This allows us to obtain more discriminative representations of the image content.
- A new classification scheme based on the dictionary response, as opposed to the reconstruction error, that allows us to exploit the representative power of the dictionaries and be robust to occlusions. This solution is naturally applicable to multi-class scenarios and preserves the local features configuration.

We experimentally validate the method we propose showing its applicability to two different classification tasks, namely single instance object recognition and object categorization. As for the first task, we use a dataset acquired in-house including 20 objects of different complexity, characterized by variability in light conditions, scale, background. In the case of categorization, instead, we consider a collection of 20 classes from the benchmark Caltech-101 dataset. In both cases, we will show that the solution we propose outperforms other approaches from the literature.

2 PRELIMINARIES

In this section we review the traditional approach to dictionary learning and describe the classification pipeline commonly used in literature in combination with such representation scheme. This will set the basis to discuss the contributions of our approach.

2.1 General Classification Framework

We first briefly introduce the classification pipeline commonly adopted with the sparse coding. It can be mainly divided in four main stages.

1. **Features Extraction.** A set of descriptors $\mathbf{x}_1, \ldots, \mathbf{x}_{m^I}$ are extracted from a test image I. Examples of local descriptors are image patches, SIFT (Lowe, 2004), or SURF (Bay et al., 2008) (either sparse or dense).
2. **Coding Stage.** The coding stage maps the input features $\mathbf{x}_1, \ldots, \mathbf{x}_{m^I}$ into a new overcomplete space $\mathbf{u}_1, \ldots, \mathbf{u}_{m^I}$.
3. **Pooling Stage.** The locality of the coded descriptors $\mathbf{u}_1, \ldots, \mathbf{u}_{m^I}$ cannot catch high level statistics of an image, therefore a pooling step is required. It can be performed at image level or with a multiscale approach (see e.g. (Boureau et al., 2010)). It has been experimentally shown that the *max pooling* operator obtains the highest performances in classification tasks (Boureau et al., 2010). With this operator an image is encoded with single feature vector $\bar{\mathbf{u}} \in \mathbb{R}^d$, where each component $\bar{u_j}$ is
$$\bar{u_j} = \max_i u_{ji} \quad \forall i = 1, \ldots, m^I \qquad (1)$$
4. **Classification** The final description is fed to a classifier such as SVM (Vapnik, 1998). Codes obtained through vector quantization usually require ad-hoc kernels to obtain good performances, instead, sparse coding approaches have shown to be effective if combined with linear classifiers, also ensuring real-time performances (Yang et al., 2009).

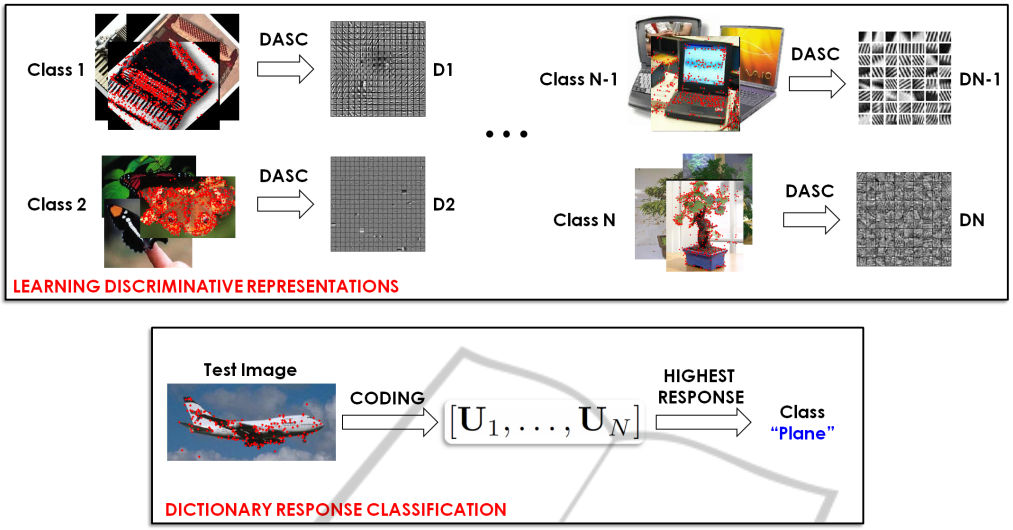

Figure 1: An overview of our framework. We consider a multi-class setting. We build a dictionary for each class proposing a new method for Discriminative and Adaptive Sparse Coding (DASC). During classification, we exploit the dictionary response rather than the reconstruction error (see text for details).

2.2 Adaptive Sparse Coding

The goal of sparse coding is to decompose a signal into a linear combination of a few elements from a given or learned dictionary. We consider the latter case where the sparse coding is adaptive, i.e. it is guided from the available data. In this case, the problem of dictionary learning may be stated as follows.

Dictionary Learning. Given a training set of images, let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_m] \in \mathbb{R}^{n \times m}$ be the matrix whose m columns $\mathbf{x}_i \in \mathbb{R}^n$ are the descriptors (feature vectors) extracted from all the images. The goal is to learn a dictionary $\mathbf{D}$ (a $n \times d$ matrix, with d the dictionary size and n the feature vector size) and a code $\mathbf{U}$ (a $d \times m$ matrix) that minimize the reconstruction error:

$$\min_{\mathbf{D},\mathbf{U}} \|\mathbf{X} - \mathbf{D}\mathbf{U}\|_F^2 + \lambda \|\mathbf{U}\|_1 \qquad (2)$$

where $\|\cdot\|_F$ is the Frobenius norm. As for the sparsity, it is known that the l_1 norm yields to sparse results while being robust to signals perturbations. Other penalties such as the l_0 norm could be employed, however the problem becomes NP-hard and there is no guarantee that greedy algorithms could reach the optimal solution.

Notice that fixing $\mathbf{U}$, the above optimization reduces to a least square problem, whilst, given $\mathbf{D}$, it is equivalent to a linear regression with the sparsifying norm $l1$. The latter problem is referred to as a feature selection problem with a known dictionary (Lee et al., 2007). One of the most efficient algorithms that converges to the optimal solution of the problem in Eq. 2 is the *feature-sign search* algorithm (Lee et al., 2007).

2.3 Classification based on the Reconstruction Error

The general classification approach described in Sec. 2.1, is appropriate for many classification tasks. However, the pooling stage, that showed to be effective especially for image categorization problems, usually loses information about spatial configuration or semantic characteristics of the features. In order to preserve these properties, which are particularly relevant e.g. for part-based object recognition, a local classification scheme is desirable. The purpose is to assign each local feature to the most likely object class. A common approach (Yang et al., 2008; Skretting and Husy, 2006; Peyré, 2009; Mairal et al., 2008a) is based on the use of the reconstruction error, defined as:

$$\mathcal{R}(\mathbf{x}, \mathbf{D}, \mathbf{u}^*) \equiv ||\mathbf{x} - \mathbf{D}\mathbf{u}^*||_F^2 \qquad (3)$$

where $\mathbf{x} \in \mathbb{R}^n$ is a feature vector, $\mathbf{D}$ is the dictionary ($n \times d$ matrix) and $\mathbf{u}^* \in \mathbb{R}^d$ is the code computed as:

$$\mathbf{u}^* = \min_{\mathbf{u}} ||\mathbf{x} - \mathbf{D}\mathbf{u}||_F^2 + \lambda \|\mathbf{u}\|_1 \qquad (4)$$

In a classification problem with N classes, if each class i is assigned a dictionary $\mathbf{D}^i$, the code $\mathbf{u}^i$ is first computed via Eq. 4 for each dictionary. Then the feature vector $\mathbf{x}$ is assigned to the class i^* that minimizes the reconstruction error $\mathcal{R}$ (Yang et al., 2008):

$$i^* = \arg\min_i \mathcal{R}(\mathbf{x}, \mathbf{D}^i, \mathbf{u}^i). \qquad (5)$$

3 LEARNING DISCRIMINATIVE REPRESENTATIONS: OUR METHOD

In this section we discuss the details of our method to learn sparse image representation and exploit it for classification tasks. Thus, recalling the general classification framework described in Sec. 2.1, we contribute to steps 2 and 4 of the pipeline. In what follows, firstly we propose a modification of the functional of Eq. 2 that also includes a contribution from the negative examples and leads to more discriminative descriptors. Then, we show that classifying each single feature using the dictionary responses rather than the reconstruction error is more effective.

3.1 Discriminative and Adaptive Sparse Coding

Unlike previous works – that do not discriminate among the different classes – our idea is to learn a suitable representation exploiting the benefit of sparsity, and introducing a further constraint on the negative examples that we force to be more densely represented. To this end, we propose a new regularized method – we called *Discriminative and Adaptive Sparse Coding* (DASC) – that we exploit to build a dictionary for each class, increasing their discriminative power.

Let us consider a supervised (classification) problem with N classes, and let $\mathbf{X}^i = [\mathbf{x}_1, \ldots, \mathbf{x}_{m^i}]$ be a $d \times m^i$ matrix whose columns are the training vectors of the $i-th$ class. Also, let $\overline{\mathbf{X}}^i = [\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^{i-1}, \mathbf{X}^{i+1}, \ldots, \mathbf{X}^N]$ be the training examples of the other classes. When learning the dictionary of class i, features belonging to it are constrained to be sparse, while features belonging to any other class j, $j \neq i$, are forced to be as smoother as possible. Hence, in order to learn the dictionary $\mathbf{D}^i$ of the $i-th$ class, which is a $n \times d^i$ matrix, where d^i is the number of atoms and n the length of the feature vector $\mathbf{x}_k$, we propose to minimize:

$$E = \begin{aligned} &||\mathbf{X}^i - \mathbf{D}^i\mathbf{U}^i||_F^2 + ||\overline{\mathbf{X}}^i - \mathbf{D}^i\overline{\mathbf{U}}^i||_F^2 + \\ &+ \lambda||\mathbf{U}^i||_1 + \mu||\overline{\mathbf{U}}^i||_2 \end{aligned} \quad (6)$$

with respect to $\mathbf{D}^i$, $\mathbf{U}^i$ and $\overline{\mathbf{U}}^i$, where $\mathbf{U}^i$ is the codes matrix of class i, while $\overline{\mathbf{U}}^i$ are the coefficients of all classes $j \neq i$. The l_2-norm induces the coefficients $\overline{\mathbf{U}}^i$ to be smooth (i.e. less sparse). So the learned dictionary still has low reconstruction error, but in addition negative examples are less sparse. As a consequence, features belonging to class i have a higher response if encoded with dictionary $\mathbf{D}^i$ rather than any other dictionary $\mathbf{D}^j$, $j \neq i$.

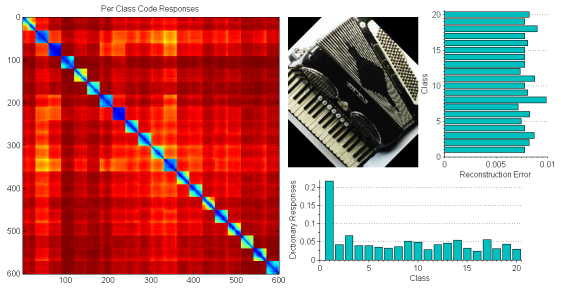

Figure 2: An intuition of the capability of the dictionaries in characterizing the corresponding class. Left: the affinity matrix. The diagonal blocks (i.e. codes of a same class) have higher similarity (blue colors) with respect to the rest (red colors). Right: a comparison of reconstruction errors and responses of the dictionaries for a test image. See text for details. Best viewed in color.

An intuition of this property is given in Fig. 2. On the left, we report the affinity matrix, obtained as the Euclidean distance among the descriptors in an all-vs-all fashion. The diagonal blocks show higher similarity, as they include codes of the same class. This speaks in favor of the capability of the dictionary of characterizing the corresponding class. On the right, we compare the reconstruction errors of a test image (in the middle of the figure) with the dictionaries response, which represents the weight of each dictionary contribution in the linear combination (see Sec. 3.3 for a formal definition). It can be easily noticed that while the reconstruction errors are comparable for all classes (even if the correct one shows a slightly lower value), the response of the correct dictionary is apparently superior than the others. This suggests us to adopt this criteria during classification: we will discuss our approach on Sec. 3.3.

3.2 Optimization Procedure

To solve Eq. 6, we apply an iterative scheme that is largely used in the dictionary learning literature. We first notice that Eq. 6 is convex in each single variable $\mathbf{D}^i, \mathbf{U}^i, \overline{\mathbf{U}}^i$ but not convex in all the variables simultaneously. The minimization of the above problem is thus carried out by clock coordinate descent (Luenberger, 2008). In particular we initialize $\mathbf{D}^i$ with d^i random examples, where d^i is the dictionary size; we fix $\mathbf{D}^i$ and $\overline{\mathbf{U}}^i$, thus we compute the solution of $\mathbf{U}^i$ via the features sign algorithm (Lee et al., 2007). Notice that minimizing Eq. 6 with respect to $\mathbf{U}^i$ is the same that minimizing Eq. 2. Indeed the other terms of the functional are constant and they do not affect the optimal solution.

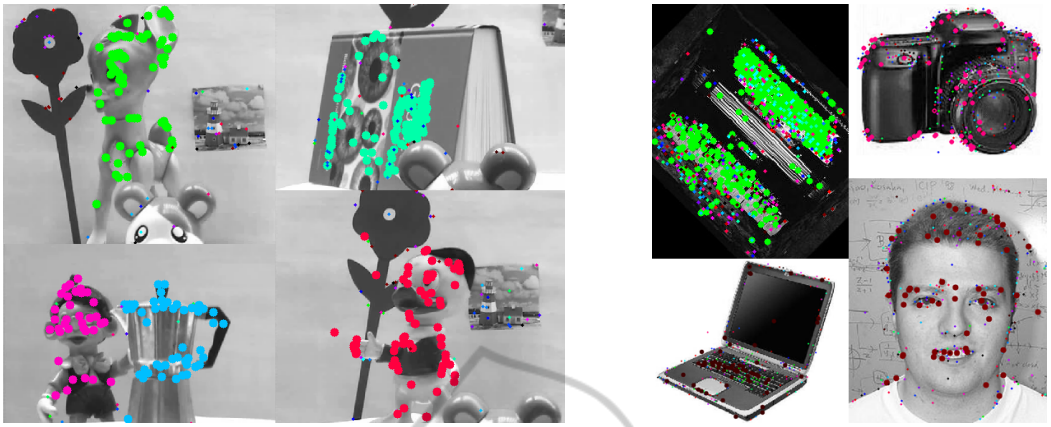

Figure 3: Examples of single features classification via dictionary response. Features are color-coded with respect to their class. Biggest circles represent features classified correctly.

Fixing $\mathbf{D}^i$ and $\mathbf{U}^i$ we can compute the solution of $\overline{\mathbf{U}}^i$ setting the gradient $\nabla_{\overline{\mathbf{U}}^i} E = 0$. For simplicity of the notation we drop the class index i:

$$\nabla_{\overline{\mathbf{U}}} E = -2\mathbf{D}^{\mathrm{T}}(\overline{\mathbf{X}} - \mathbf{D}\overline{\mathbf{U}}) + 2\mu\overline{\mathbf{U}} \quad (7)$$

Setting the derivative to zero we obtain:

$$\nabla_{\overline{\mathbf{U}}} E = 0 \Rightarrow \overline{\mathbf{U}} = (\mathbf{D}^{\mathrm{T}}\mathbf{D} + \mu\mathbf{I})^{-1}\mathbf{D}^{\mathrm{T}}\overline{\mathbf{X}} \quad (8)$$

where $\mathbf{I}$ is the identity matrix $d^i \times d^i$. Finally, fixing $\mathbf{U}$ and $\overline{\mathbf{U}}$ we update the solution of $\mathbf{D}$ as:

$$\nabla_{\mathbf{D}} E = -2(\mathbf{X} - \mathbf{D}\mathbf{U})\mathbf{U}^{\mathrm{T}} - 2(\overline{\mathbf{X}} - \mathbf{D}\overline{\mathbf{U}})\overline{\mathbf{U}}^{\mathrm{T}} \quad (9)$$

and setting it to zero $\nabla_{\mathbf{D}} E = 0$ we obtain:

$$\mathbf{D} = (\mathbf{X}\mathbf{U}^{\mathrm{T}} + \overline{\mathbf{X}}\overline{\mathbf{U}}^{\mathrm{T}})(\mathbf{U}\mathbf{U}^{\mathrm{T}} + \overline{\mathbf{U}}\overline{\mathbf{U}}^{\mathrm{T}})^{-1} \quad (10)$$

This optimization process is repeated for a fixed number of iterations.

3.3 Classification based on Dictionary Response

As discussed on Sec. 3.1, the classification based on the reconstruction error does not guarantee the correctness of the results.
In literature several experiments confirmed that the max pooling operator achieves the best results in terms of classification (Boureau et al., 2010). This means that dictionary atoms with higher responses are more representative of the underlying distribution of the data. Inspired by this consideration and experimentally observing the higher discriminative power of dictionary responses, we propose a classification method based on the intensity of dictionary responses after the coding stage.

We start by considering the *Global Dictionary* $\mathbf{D} = [\mathbf{D}^1, \dots \mathbf{D}^N]$, with $d = \sum_{i=1}^{N} d^i$ atoms, as the concatenation of all the class dictionaries previously learned. We recall that a signal $\mathbf{x} \in \mathbb{R}^n$ can be decomposed into a linear combination of dictionary and codes, i.e. $\mathbf{x} = \mathbf{D}\mathbf{U}$, with $\mathbf{U}$ a $d \times 1$ column vector. Therefore we can interpret the code $\mathbf{U}$ as the relevance of each dictionary atom in the linear combination. Assuming to know, as in our framework, which atoms of the dictionary describe a certain class, $\mathbf{U}$ can be interpreted as a concatenation of blocks, each one including the responses of a dictionary:

$$\mathbf{U}^{\mathrm{T}} = [\mathbf{u}^1, \dots, \mathbf{u}^N]; \quad (11)$$

where $\mathbf{u}^i$ is a vector of size d^i representing the response of the i-th dictionary. We can define the response H of the i-th class as:

$$H(\mathbf{u}^i) = \sum_{j=1}^{d^i} u_j^i \quad (12)$$

where $\mathbf{u}^i$ is the code corresponding to the i-th block of the global dictionary $\mathbf{D}$ and d^i is the size of the class dictionary. At this point we can assign each local feature $\mathbf{x}$ to the class i such that:

$$i^\star = \arg\max_i H(\mathbf{u}^i) \quad (13)$$

This classification scheme has many advantages: first it exploits the dictionary learning method previously described, indeed dictionaries will have higher response for features belonging to their class. Secondly it preserves the local information of the features: infact no pooling stages are required, therefore we do not lose the spatial configuration of the descriptors. Finally it is natural its application in a multi-class classification scenario: indeed an image will be classified using the highest sum of all the local responses.

Figure 3 shows different examples of features classification in which higher dictionary responses are

Figure 4: The dataset we used for experiments on single instance object recognition. It has been acquired in-house and includes 20 objects of different complexity.

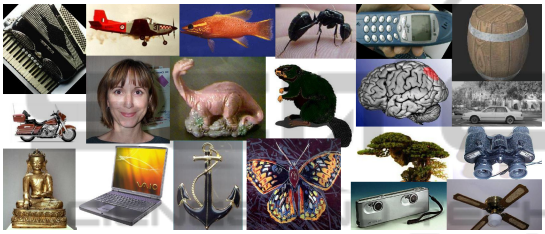

Figure 5: A selection of 20 classes from the popular Caltech-101 dataset, that we considered within the object categorization experiments.

highlighted with bigger circles. It is apparent that although the presence of misclassified elements, clusters of coherent features can be detected which reveal the presence of known objects.

4 EXPERIMENTS

In this section we experimentally validate the proposed method for application to object classification. In particular we consider two different problems, namely *Single Instance Object Recognition* and *Object Categorization*. In the first class of problems, we want to classify a specific instance of an object, while within categorization tasks many different instances of an object class must be recognized. To this end, we consider two different datasets, with characteristics appropriate for the two experimental scenarios. For the first one, we used a dataset acquired in-house, composed of 20 objects (see Fig. 4) of different complexities. It includes both planar objects (as books and boxes) and objects with a more complex 3D structure (as puppets). It represents an appropriate testbed for our purposes thanks to the objects variability as well as the availability of a significant number of samples. In fact, for each object a training video of approximately 300 frames is provided which is acquired with a still camera that observes the objects as they rotate, thus including images from different view-points. Also, test videos with different characteristics (as variable background, light, scale) are provided. Although the availability of video, no temporal information is exploited in our approach, where images are processed independently. We used 30 images per class as training set, while the remaining frames have been used as test set.

For what concerns object categorization, we considered a selection of classes from the very popular *Caltech-101* dataset (Fei-fei et al., 2006). In this case, several instances of an object class are provided. The complexities of this dataset reside in object variability, cluttered background and presence of occlusions. In our experiments we used a subset of 20 classes (see Fig. 5) for computational reasons (we recall that we need to build a unique dictionary matrix, that may have very high dimensions). For each class we used 30 of the available images as training set, while the others have been used for the test phase (max 50 per class).

The structure of our method does not depend on the type of features extracted from the images. In our experiments,we first run a corner detector and then compute SIFT descriptors (Lowe, 2004).

4.1 Analysis of the Dictionaries

We first quantitatively evaluated the learned dictionaries in terms of reconstruction error and nonzero elements with respect to the level of sparsity of the obtained representation. We extracted the feature vectors from each image, and then we learned the dictionary of each class (object instance in single instance object recognition, category in object categorization) accordingly to the procedure described on Sec. 3. Recalling the notation of Sec. 3, we used $m^i = 1000$ features for each class and fixed the dictionary size to $d^i = 512$.

The parameter $\mu = 0.15$ of Eq. 6 has been selected with a cross-validation procedure on the reconstruction error. Finally, we coded a test set of descriptors using Eq. 4. The results show that descriptors belonging to a given class i obtain a lower reconstruction error when using the dictionary blocks corresponding to $\mathbf{D}^i$ than the others. For what concerns the percentage of nonzero elements, analogously, those features obtained the highest response from the corresponding dictionary.

We report in Fig. 6 examples of two classes of different complexity from the Caltech-101. The plots show the trends of reconstruction error (first row) and the percentage of nonzero elements (second row) as

the parameter lambda (i.e. the one controlling the sparsity) increases. It is apparent that the best performing dictionary is the correct one.

To measure the goodness of the dictionaries, we compute average reconstruction error and dictionary response on the two datasets by selecting the values corresponding to a reference $\lambda = 0.15$. We first evaluated the average reconstruction error for all the test images when using the correct dictionary or another one, obtaining respectively 6×10^{-4} and 7×10^{-3} for the single instance dataset, and about 3×10^{-3} for both in the case of Caltech-101. Similarly, we computed the average dictionary responses evaluating the weight of the correct dictionary as opposed to the others in the sum in Eq. 12. This measures not only the number of nonzero elements per dictionary, but also actual contribution of the codes in the linear combination. In this case we obtained that the weight of the codes of the correct dictionary is on average the 53.41% for the single instance dataset, and the 26.21% for the Caltech-101. On the contrary, the weight of the other dictionaries is on average the 2.33% per dictionary in the case of single instance object recognition, and the 3.78% per dictionary of the Caltech-101.

4.2 Classification Results

Again, we conduct this experiment in both single instance and categorization problems. Following a common procedure, we use a K-fold validation strategy for parameters tuning. We consider $K = 10$ different runs and randomly select training and test sets, to obtain a reliable statistic. The average per-class recognition rates were stored at each run. We report as final results the recognition rates averaged over the runs.

For both single instance object recognition and object categorization, we compare the performances of classification based on the reconstruction error (Sec. 2.3) with the approach we propose (Sec. 3.3), based on the evaluation of the dictionary responses, both coupled with our functional (Eq. 6). Also, on top of the proposed learned coded we consider a linear classifier comparing the approach described in (Yang et al., 2009), which does not consider prior information on the classes, as opposed to the use of the dictionaries of each class trained as described in Sec. 3.

Table 1 reports the obtained results. The proposed classification scheme based on the dictionary responses outperforms the one based on the reconstruction error, but it is still far from comparing the accuracy obtained by employing a learning algorithm such as SVM. SVMs with per class dictionaries obtains better results with respect to the traditional approach (Yang et al., 2009).

Table 1: Accuracy results for single instance object recognition and object categorization.

Method	S. I. Obj. Rec.	Obj. Cat.
Reconstruction Error	78.41%	22.23%
Dictionary Response	**89.32%**	**59.68%**
SVM + Dictionary	94.12%	76.95%
SVM + Class Dictionary	**97.21%**	**84.43%**

5 DISCUSSION

In this work we tackled the problem of finding compact and discriminative image representations by means of the sparse coding theory. We considered a multi-class classification setting typical of object recognition and image categorization. We proposed to modify the standard dictionary learning functional by adding a term accounting for the negative samples and forcing them to be associated with smoother and denser descriptors. On the contrary, positive samples descriptors are constrained to be sparse. We showed that this approach provides highly discriminative representations and is very effective from the computational standpoint thanks to compactness and usability with linear kernels.

We also showed that the dictionary responses can be directly used as a criteria for image feature classification in a new scheme we proposed here. Thanks to this formulation, important features properties, as e.g. the spatial configurations, can be kept and exploited for subsequent steps in the analysis (as in part-based object recognition).

To show the generality of our approach, we considered object classification from two different perspectives. In particular, we faced the problem of single instance object recognition and object categorization. We used two different dataset that captured the peculiarities of the two scenarios, namely an in-house acquired dataset of 20 objects for the first problem, and a selection of 20 classes from the well-known Caltech-101 for the latter. The experimental results spoke in favor of our approach, that performs better than other commonly adopted solutions and showed that the classification based on dictionary responses is more effective than the one based on the reconstruction error. Also, classifying single features allowed us to cope with cluttered background and occlusions among objects. Adding a final classification step, as an SVM, further improves the recognition rates, even if the spatial information of objects are lost due to the pooling operator.

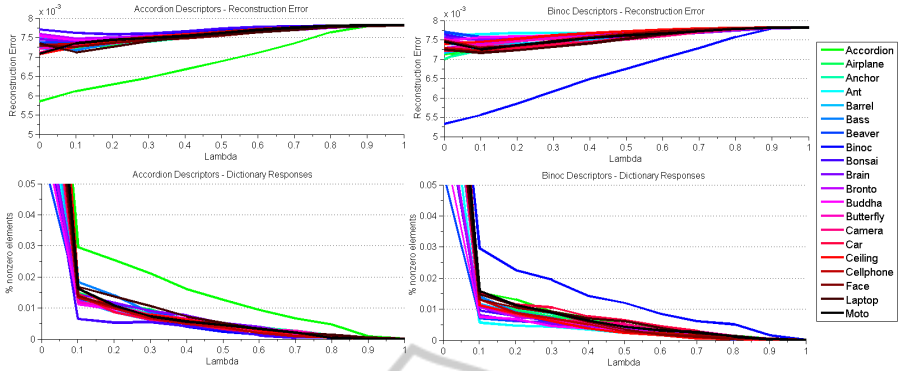

Figure 6: Examples of reconstruction error and percentage of nonzero codes for two objects from the Caltech-101 dataset (object categorization task). Both the trends of the reconstruction error (above) and of the percentage of nonzero elements (below) as the sparsity parameter increases show that the best performing alphabet is the correct one.

Future extensions of this work will consider the design and the development of a method for part-based object detection and recognition built on top of our current achievements. The idea we are pursuing is based on building part related dictionaries and exploiting the dictionary response classification scheme for detection purposes and the temporal information. Our final goal is to overcome the common *sliding window* approach for object localization, with a more efficient part-based localization method.

REFERENCES

Bay, H., Ess, A., Tuytelaars, T., and Vangool, L. (2008). Speeded-up robust features. *CVIU*, 110:346–359.

Boureau, Y.-L., Bach, F., LeCun, Y., and Ponce, J. (2010). Learning mid-level features for recognition. In *CVPR*.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *CVPR*.

Destrero, A., De Mol, C., Odone, F., and A., V. (2009). A sparsity-enforcing method for learning face features. *IP*, 18:188–201.

Fei-fei, L., Fergus, R., and Perona, P. (2006). One-shot learning of object categories. *PAMI*, 28:594–611.

Hasler, S., Wersing, H., Kirstein, S., and Körner, E. (2009). Large-scale real-time object identification based on analytic features. In *ICANN*.

Hasler, S., Wersing, H., and Krner, E. (2007). A comparison of features in parts-based object recognition hierarchies. *ICANN*.

Jia, Y., Huang, C., and Darrel, T. (2012). Beyond spatial pyramids: Receptive field learning for pooled image features. In *CVPR*.

Lee, H., Battle, A., Raina, R., and Ng, A. Y. (2007). Efficient sparse coding algorithms. In *NIPS*.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *IJCV*, 60:91–110.

Luenberger, D. G. (2008). *Linear and Nonlinear Programming*. Springer.

Mairal, J., Bach, F., and Ponce, J. (2012). Task-driven dictionary learning. *PAMI*, 34:791–804.

Mairal, J., Bach, F., Ponce, J., Sapiro, G., and Zisserman, A. (2008a). Discriminative learned dictionaries for local image analysis. In *CVPR*.

Mairal, J., Bach, F., Ponce, J., Sapiro, G., and Zisserman, A. (2008b). Supervised dictionary learning. In *NIPS*.

Olshausen, B. A. and Fieldt, D. J. (1997). Sparse coding with an overcomplete basis set: a strategy employed by v1. *Vision Research*.

Peyré, G. (2009). Sparse modeling of textures. *Journal of Mathematical Imaging and Vision*, pages 17–31.

Skretting, K. and Husy, J. (2006). Texture classification using sparse frame based representation. *EURASIP Journal on Applied Signal Processsing*.

Vapnik, V. (1998). *Statistical Learning Theory*. John Wiley and Sons, Inc.

Viola, P. and Jones, M. (2004). Robust real-time face detection. *IJCV*, 57:137–154.

Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., and Gong, Y. (2010). Locality-constrained linear coding for image classification. In *CVPR*.

Wersing, H. and Körner, E. (2003). Learning optimized features for hierarchical models of invariant object recognition. *Neural Computation*.

Yang, J., Wright, J., Ma, Y., and Sastry, S. (2008). Feature selection in face recognition: A sparse representation perspective. *PAMI*.

Yang, J., Yu, K., Gong, Y., and Huang, T. (2009). Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*.

Yang, J., Yu, K., and Huang, T. (2010). Efficient highly over-complete sparse coding using a mixture model. In *ECCV*.