

**SPARSE REPRESENTATIONS FOR IMAGE CLASSIFICATION:
LEARNING DISCRIMINATIVE AND
RECONSTRUCTIVE NON-PARAMETRIC DICTIONARIES**

By

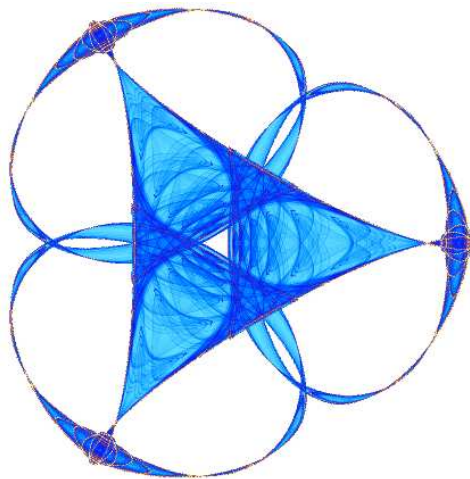
Fernando Rodriguez

and

Guillermo Sapiro

IMA Preprint Series # 2213

(June 2008)



INSTITUTE FOR MATHEMATICS AND ITS APPLICATIONS

UNIVERSITY OF MINNESOTA
400 Lind Hall
207 Church Street S.E.
Minneapolis, Minnesota 55455-0436

Phone: 612-624-6066 Fax: 612-626-7370

URL: <http://www.ima.umn.edu>

Sparse Representations for Image Classification: Learning Discriminative and Reconstructive Non-Parametric Dictionaries

Fernando Rodriguez and Guillermo Sapiro

University of Minnesota

December 2007

Abstract. A framework for learning optimal dictionaries for simultaneous sparse signal representation and robust class classification is introduced in this paper. This problem for dictionary learning is solved by a class-dependent supervised simultaneous orthogonal matching pursuit, which learns the intra-class structure while increasing the inter-class discrimination, interleaved with an efficient dictionary update obtained via singular value decomposition. This framework addresses for the first time the explicit incorporation of both reconstruction and discrimination terms in the non-parametric dictionary learning and sparse coding energy. The work contributes to the understanding of the importance of learned sparse representations for signal classification, showing the relevance of learning discriminative and at the same time reconstructive dictionaries in order to achieve accurate and robust classification. The presentation of the underlying theory is complemented with examples with the standard MNIST and Caltech datasets, and results on the use of the sparse representation obtained from the learned dictionaries as local patch descriptors, replacing commonly used experimental ones.

1 Introduction

The study of sparse representations has become a major field of research in signal processing. Efforts have been focused mainly on the development of theoretical frameworks (e.g., [2, 5]), algorithms to efficiently perform sparse coding (e.g., [4, 8, 21]), learning of overcomplete sets of vectors denoted as *dictionaries* (e.g., [1, 23, 28]), and applications in image processing (e.g., [6, 20, 28]). Sparse representations over non-parametric learned dictionaries lead to state-of-the-art results for image enhancement [20].

Since originally trained to contain sufficient information for reconstruction, sparse representations are, from the point of view of signal classification, a *reconstructive* approach. This provides representations that are relatively robust against distortions and missing data. On the other hand, *discriminative* methods have as criteria the classification performance itself, an element that has not been significantly addressed yet in the sparsity (non-parametric) dictionary learning community and that constitutes one of our key contributions in this work. Discriminative methods often outperform reconstructive ones in ideal conditions, but lack robustness.

Our proposed framework introduces a novel metric which includes both reconstruction and discrimination terms in the dictionary learning process, benefitting from the best of both discriminative and reconstructive worlds. This is incorporated into a new energy, inspired by the framework put forward in [1, 20], leading to the learning of adapted dictionaries and sparse discriminative and reconstructive image representations with them. These learned dictionaries provide robust discriminant representations through adaptation to the dataset. Our proposed framework is based on the concept of obtaining simultaneous sparse decompositions within each class, so as to extract its internal structure, while keeping a global discrimination term among different classes. Such explicit and efficient incorporation of the task (classification) into the dictionary learning for sparse coding is unique and a key novelty of the proposed work.

1.1 Related Work and Our Contribution

Huang and Aviyente, [11], proposed the marriage of discrimination and reconstruction in sparse image representations, introducing a novel discrimination term into the classical reconstructive energy formulation of sparse coding. Their approach proved to yield robust and discriminant image representations through an intrinsic dimensionality reduction. In contrast with our proposed framework, there is no dictionary learning in [11], and they use pre-defined dictionaries and sparse coding over them. As shown below, and it is further supported by the image processing literature on non-parametric dictionary learning, adapted learned dictionaries outperform off-the-shelf ones. Effrosyni and Frossard introduced a similar algorithm [14], which they named *Supervised Simultaneous Orthogonal Matching Pursuit* or SSOMP. Simultaneous sparse decompositions, which are applied to the whole dataset/class at once, are proven to be essential in order to extract the structure of a class and to help capturing its intrinsic variability.

LeCun *et al.* introduced an algorithm for learning sparse representations, based on a energy model, through a linear coder and a linear decoder [26, 27]. This is based on a (coordinate) sparsifying logic quite different from our principle of sparse coding. The work includes a complex neural network, with multiple layers and training steps. The dictionary design and neural network training are based on different criteria, it is not clear how much of the outstanding performance is due to each part. No results are given concerning robustness, which are needed to verify whether the properties from sparse coding have been inherited. Our objective is learning representations that by themselves are discriminative and robust, leading both to different energy formulations and optimization techniques. We include explicit discriminative terms, which are absent in the framework in [26, 27], in addition to the reconstructive one.

Lazebnik and Raginsky recently introduced an elegant dictionary learning algorithm based on *Information Loss Minimization* [16]. It learns a codebook with the objective of obtaining a quantization that does not cause high distortion and at the same time keeps nearly all the information about the class of the original signal. There is no explicit discrimination constraints, although classification is the main goal. Leibe and collaborators, e.g., [7, 17], have proposed in a series of leading works the use of learned dictionaries, these obtained from clustering of

image patches, without explicit sparsity, reconstruction, and/or discrimination requirements. Sparsity also has been recently incorporated into a very interesting robust face recognition framework by Ma *et al.*, [32], motivated by the work on compressed sensing and random projections. This work does not explicitly enforce reconstruction and/or discrimination neither it learns adapted dictionaries. In [9] and companion papers, the authors develop an efficient l_1 -based optimization approach for learning dictionaries for sparse representation, much of the spirit of [1], again with an energy tuned to reconstruction only, and use the coefficients of the overcomplete representation for classification. The minimal reconstruction error from multiple generative-only dictionaries, each one independently learned for a different class, is used in [25] for classifying textures. Finally, we should mention that the work on epitomes, [12], provides a different generative-only model for learning dictionaries that can also be used for recognition [15].

Contributions: In contrast with previous approaches, the framework here proposed learns a non-parametric dictionary which is efficient for sparsely representing a signal and at the same time performing class discrimination. Such dictionary and sparse representation are derived from the efficient minimization of an energy that explicitly includes these critical components. This novel classification framework can be seen as a deviation and step forward from approaches that either use off-the-shelf dictionaries and features (e.g., [11, 18, 22, 31]), or learn dictionaries without explicit discrimination and/or reconstruction goals (meaning they obtain or learn dictionaries with a criteria that often does not explicitly include the actual application and performance criteria). Although such alternative approaches have performed outstandingly, their actual optimality, performance, and limitation studies have been purely experimental. The underlying idea behind our proposed framework is to start gearing toward the design of feature detectors and non-parametric dictionaries that are designed and optimized for the task at hand.

2 Learning Dictionaries for Representation and Discrimination

We now build, step-by-step, the proposed framework for learning discriminative and representative non-parametric dictionaries.

2.1 Supervised Sparse Coding

For sparse coding, we will extend, by adding discrimination power, the *Simultaneous Orthogonal Matching Pursuit (SOMP)* algorithm, see [24, 29, 30] for details and theoretical results on this greedy technique. Given a dictionary matrix $D \in \mathcal{R}^{n \times K}$ (which we will later learn), that contains K atoms $\{\mathbf{d}_j\}_{j=1}^K \in \mathcal{R}^n$ ($K \geq n$), and a set of signals $\{\mathbf{x}_j\}_{j=1}^s \in \mathcal{R}^n$, SOMP attempts to represent these signals *at once* as a linear combination of a *common* subset of atoms of cardinality much smaller than n (sparse representation). Under the assumption that those signals belong to a certain class, SOMP attempts to extract their common

internal structure. By keeping the sparsity low enough, the internal variation of the class could be eliminated, leading to more accurate classification while being robust to noise. After adding classification terms into SOMP, see next, we will explicitly use these coefficients of sparse representation, over a discriminative learned dictionary, for classification.

To further increase the inherent discriminant capacity of SOMP, we will next incorporate into SOMP a discrimination measure inspired by *linear discriminant analysis* (LDA) (on top of the original reconstruction component, see also [11, 14]), the quotient of the l_2 norms of the corresponding scatter matrices. Given c sets of vectors, each one representing one class, $\{\alpha_i^j\}_{i=1}^{n_j}$ with $j \in \{1, \dots, c\}$ and $\alpha_i^j \in \mathbf{R}^K$, we propose as linear discrimination measure $J\left(\{\{\alpha_i^j\}_{i=1}^{n_j}\}_{j=1}^c\right) := \frac{\text{trace}(\mathbf{S}_B)}{\text{trace}(\mathbf{S}_W) + \mu}$, where \mathbf{S}_B and \mathbf{S}_W are the standard *between-classes* and *within-class* scatter matrices and μ is a regularization parameter.¹

Following the introduction of the discriminative measure $J(\cdot)$, the originally purely reconstructive objective of SOMP is modified incorporating this discrimination measure over the sparse representation coefficients $\{\{\alpha_i^j\}_{i=1}^{n_j}\}_{j=1}^c$ corresponding to each one of the c classes. For a dictionary $\{\mathbf{d}_j\}_{j=1}^K$ and a set of indices Λ , let $\Phi_\Lambda \in \mathcal{R}^{n \times |\Lambda|}$ be the matrix whose columns are the \mathbf{d}_i , $i \in \Lambda$. Let $\{\mathbf{x}_i^j\}_{i=1}^{n_j}$ be the signals from the j -th class, $j \in \{1, \dots, c\}$ (e.g., images in column representation), and \mathbf{X} the matrix with columns $\{\{\mathbf{x}_i^j\}_{i=1}^{n_j}\}_{j=1}^c$. We state the *Simultaneous Sparse Discriminant Problem* as

$$\max_{\Lambda, |\Lambda| \leq L} \left\{ \theta \cdot \underbrace{J((\Phi_\Lambda^T \Phi_\Lambda)^{-1} \Phi_\Lambda^T \mathbf{X})}_{\aleph} - \underbrace{\|\mathbf{X} - \Phi_\Lambda (\Phi_\Lambda^T \Phi_\Lambda)^{-1} \Phi_\Lambda^T \mathbf{X}\|_F^2}_{\aleph} \right\}. \quad (1)$$

Here, L is the sparsity factor indicating how many atoms are used to represent the signals, \aleph is simply the orthogonal projection of the signal onto the selected set Φ_Λ , and θ is a parameter that controls the trade-off between the discriminative term (first component of (1)) and the reconstruction term (second component of (1), where $\|\cdot\|_F$ stands for the Frobenius norm, and only one present in the original SOMP). θ is dynamically updated (see also [14]).² We propose a greedy approach to address this optimization, denoted as *Supervised SOMP* (*SSOMP*), see Figure 1 (in the following we omit the dynamic dependency of θ in the notation).

Considering that the sparsity coefficients are $\alpha_i^j = (\Phi_\Lambda^T \Phi_\Lambda)^{-1} \Phi_\Lambda^T \mathbf{x}_i^j$, the corresponding scatter matrices of $\{\{\alpha_i^j\}_{i=1}^{n_j}\}_{j=1}^c$ and $\{\{\mathbf{x}_i^j\}_{i=1}^{n_j}\}_{j=1}^c$ verify $\mathbf{S}_A(\alpha) = (\Phi_\Lambda^T \Phi_\Lambda)^{-1} \Phi_\Lambda^T \mathbf{S}_A(\mathbf{x}) \Phi_\Lambda (\Phi_\Lambda^T \Phi_\Lambda)^{-1}$, where $A \in \{B, W\}$. Under the assumption that the degree of correlation is low, $(\Phi_\Lambda^T \Phi_\Lambda)^{-1}$ can be approximated by the identity and thus we decompose the contribution of each vector,

¹ Ideally, we would like to use the product of the positive eigenvectors of those matrices.

Since the determinant is zero ($c < n$), it is not possible to use it. We then chose the summation to yield our discrimination term.

² $\theta^{(t)} \leftarrow \theta \cdot \frac{1}{K} \sum_{j=1}^c \sum_{i=1}^{n_j} \sum_{p=1}^K | \langle \mathbf{r}_i^{j,(t-1)}, \mathbf{d}_p \rangle |$, with $\mathbf{r}_i^{j,(t-1)}$ being the previous residual, Figure 1.

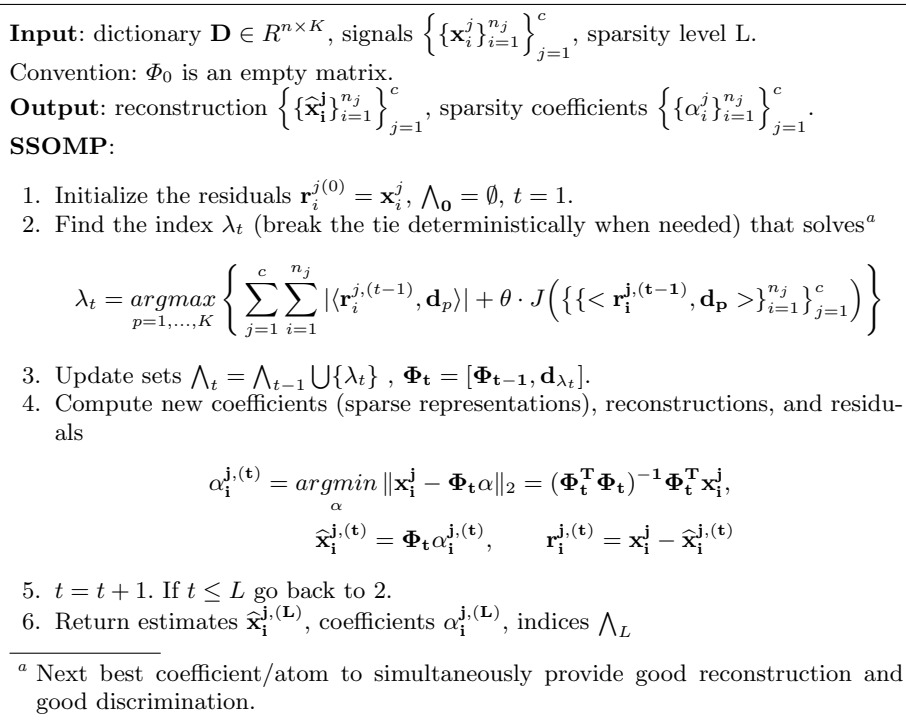


Fig. 1. Supervised SOMP (SSOMP).

$\operatorname{trace}(\mathbf{S}_A(\alpha)) \approx \sum_{i=1}^L \mathbf{d}_{\lambda_i}^T \mathbf{S}_A(\mathbf{x}) \mathbf{d}_{\lambda_i}$, where, as in Figure 1, $\Lambda = \{\lambda_1, \dots, \lambda_L\}$. This quantity can be greedily calculated. Furthermore, to yield a better estimate, we evaluate each one of the summation terms in this expression over the residuals, so that the non orthogonality is better taken into account. This is equivalent to the way that classical OMP treats correlations in the orthogonal projection to evaluate the reconstruction error. Furthermore, it is equivalent to applying a one-dimension dimensionality reduction over the residual, so that we can directly use J , as stated in the algorithm. Finally, note that there is no need to build any matrix explicitly, since it can be directly evaluated.

2.2 Class Supervised SOMP

The SOMP extracts the common coherent internal structure of a given class. However, when dealing with signals from different classes, this coherence does not exist any more. Thus the number of atoms needed to give a proper characterization of all the classes with SOMP is larger than that needed for each class individually. As a consequence, the representation captures more of the intra-class variance, decreasing the classification performance. Performing one independent SOMP per class, and joining the sets of atoms selected could be even worse, since there may well be a minimum common structure among the classes and redundancies could arise, giving rise to problems such as multiple representations. This fact is critical when the atoms have been trained for re-

construction tasks, since a small number of them can highly accurately describe the signals.

In order to achieve the goal of seeking internal structure within each class and at the same time global discrimination among the classes, we propose the *Class Supervised SOMP (CSSOMP)* algorithm in Figure 2. The reconstruction term is treated class per class, whereas the discrimination one is always global.

Now that we know how to sparsely encode signals to simultaneously achieve discrimination and reconstruction (thereby robustness), it is time to optimize the dictionary to the data and task, bringing an additional novelty to the framework.

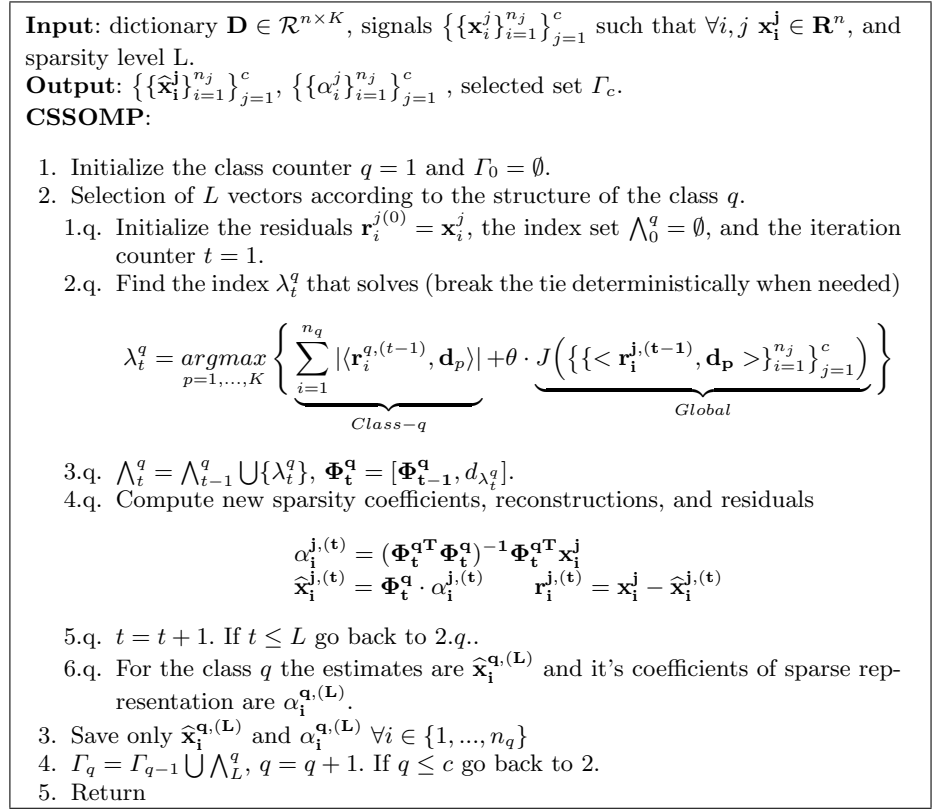


Fig. 2. *Class Supervised SOMP (CSSOMP).*

2.3 Learning the Dictionary: The Complete Model

In order to learn dictionaries that are also discriminant, we define the *Sparse Discriminant Dictionary Problem*:

$$\max_{\mathbf{D}, \alpha} \left\{ \theta \cdot J\left(\{\{\alpha_i^j\}_{i=1}^{n_j}\}_{j=1}^c\right) - \sum_{j=1}^c \sum_{i=1}^{n_j} \|\mathbf{x}_i^j - \mathbf{D}\alpha_i^j\|_2^2 \right\}, \quad (2)$$

subject to $\|\alpha_i^j\|_0 \leq L, \forall i, j$. In contrast with the energies in previous sections, the optimization is both over the dictionary \mathbf{D} and the sparse representation over it, α . If $\theta = 0$, we obtain the reconstruction only formulation in [1].

To address this optimization problem, we extend the K-SVD, an algorithm for learning overcomplete non-parametric dictionaries for sparse representation [1]. Its objective is to design a dictionary such that the reconstruction error over a set of signals, when coded sparsely, is minimal. This is achieved through an iterative process which alternates a *Sparse Coding* stage which follows the classical OMP, and a *Dictionary Update* stage derived from simple SVD (each atom is updated to improve the reconstruction of those signals that use it). Our proposed algorithm modifies the Sparse Coding Stage, which is now performed by a CSSOMP instead of OMP, adding the discrimination component, and obtaining the *Supervised K-SVD (SKSVD)*, Figure 3.

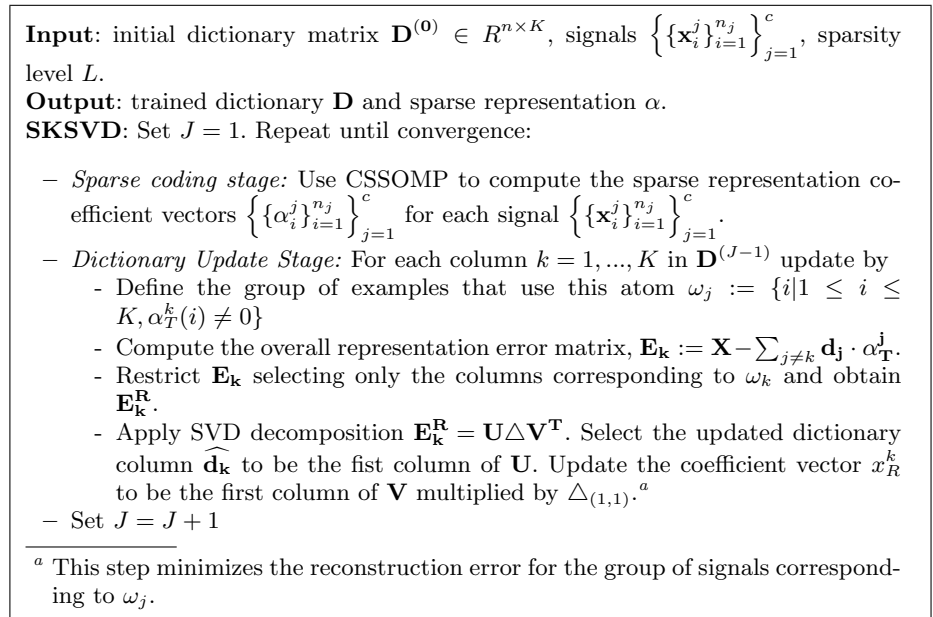


Fig. 3. *Supervised K-SVD (SKSVD).*

The proposed CSSOMP permits to maximize the energy according to all of the α_i^j , keeping in mind the global aspect of the discrimination term. At the same time we incorporate the prior of a coherent structure in a class. This is not only obtained through the simultaneous decomposition of each class, but also from the transmission of the information between both coding and dictionary update stages. All signals from the same class will use the same atoms, and thus the Dictionary Update Stage accounts for the general internal structure of the class (specially if the sparsity is kept small compared to the dimension).³

³ Explicitly incorporating an additional discrimination term into the dictionary update step is the subject of parallel efforts, [19].

Since the proposed framework permits the use of common atoms by multiple classes, the inner common structures of the whole ensemble of atoms will be learned by those atoms used by multiple classes. Finally, compared to selecting a set of atoms simply through CSSOMP, the learning will adapt to any desired amount of atoms, being fixed at first, so that we force during the learning stage the use of atoms by multiple classes. We have thereby eliminated the problem of divergence of supports.

3 Experimental Results

Experimental analysis to demonstrate the importance of learning discriminative and representative (non-parametric) dictionaries has been first carried out with the standard MNIST Handwritten Digit Database, $n = 16 \times 16 = 256$ dimensional vectors. This is done to demonstrate the importance of the proposed framework, and in particular of learning non-parametric dictionaries that reconstruct and discriminate. We then show results on natural images.

The classification tasks have been performed using linear SVMs on the sparse representation coefficients (see also [9]), coherently with our criteria. In particular we have used the implementation in [3] and a multiclass one-against-one strategy [10]. Following [13], and given the large amount of data, parameters and accuracies have been estimated through 10-fold Cross-Validation. SVMs are trained in noiseless conditions and tested over corrupted data. Then, the robustness comes from the representation itself.⁴

The particular dictionaries used in the test are: (1) union of DCT and Haar basis (511 total atoms), which are very well adapted to this database; (2) KSVD learned dictionary of size 1031 and sparsity $L = 4$ (no discrimination components); and (3) the introduced SKSVD learned for dimension (total number of atoms) 50 and sparsity $L = 15$, yielding an under-complete system. Dictionaries are learned with 20000 images and the different algorithms are tested with about 9000 images.

The Need for Simultaneous Decompositions- The first idea we could have is to directly use the coefficients of a sparse decomposition through OMP for classification, under the hypothesis that different classes will have different supports. Unfortunately, this does not hold. First of all, a sparse representation over over-complete dictionaries has the problem of multiple representations. Through the SOMP we reduce the size of the dictionaries (to 50), thereby addressing this first problem. Even then, for all of the dictionaries, the distribution of Hamming distances among supports (α -s) is very similar within one class and between classes. Even worse, the average distance is close to the maximum possible. Learning the dictionary with KSVD, and in particular with the proposed SKSVD, yields

⁴ There is a link between SVMs feature selection techniques and our discrimination term. This could be interpreted as incorporating the F-Score criteria itself in the design of the dictionary. However, those techniques do not take into account possible correlations between variables and reconstruction properties, SVMs is a discriminative approach.

significantly better results than the fixed dictionary one, but this is still not sufficient for classification.

The underlying problems are the non ideality of OMP and the over-completeness of our dictionaries. Consider the images for the class corresponding to the number “1.” There may well be 15 atoms of the KSVD dictionary that describe the ensemble accurately and show the internal coherence within the class. However, when one performs an OMP decomposition for a single number “1,” the over-completeness of the dictionaries and their good reconstruction capacity may well drive the greedy selection towards different atoms tailored specifically for it. This is illustrated in Figure 4 for the DCT+Haar dictionary. In the case of OMP, after 5 atoms, the algorithm selects highly localized Haar atoms, in order to describe small details. The reconstructions are not natural and focus on certain areas of the images. Details are important for reconstruction, but not necessarily for classification, since they are often associated to the intra-class variation. However, in the case of SOMP over two classes together (Figure 4-center), and each one of them separately (Figure 4-bottom), all the atoms are dedicated to the general shape, mostly DCTs. The reconstructions are blurred, with no details, but keep the essential structure of the number/image (class). They achieve the extraction of the common internal structure. This is why simultaneous decomposition is so important in classification. The difference between SOMP global and per-class is not very high in this particular example since there are only two classes. Nevertheless, it is remarkable that for the per-class over “1”-s, instead of selecting as first atom the DC one, it selects a DCT that has the shape of a vertical stroke.

The sparsity has to be kept small not to capture those intra-class variations associated to details and, at the same time, the maximal number of atoms that could be selected is $c \cdot L$. Previous arguments show that as a matter of fact the number should be much smaller than this quantity. Since multiple representation is a major concern, we train highly under-complete dictionaries. Once those atoms (dictionary) are selected, multiple characterizations could then be envisioned, such as correlation with those atoms, OMP with fixed error instead of sparsity, and OMP with fixed sparsity. Even when the set is under-complete, the problem with OMP remains and correlation does not do better. The right descriptor are then the coefficients of orthogonal projection over the span of those atoms.⁵ In fact, our criteria treats the intra-class variations as *noise*, trying not to capture them in the selected subset of atoms. Orthogonal projection follows this objective and unifies the representation.

Robustness of the Framework- In order to study the robustness of our approach and to compare it with the fixed dictionary, we test the classification performance, for the MNIST data, under noisy (additive Gaussian noise) and random occlusion conditions. For each dictionary we obtain 9 different representations (for the different types of dimensionality reduction), with dictionary sizes of 15, 30, and 50, both for the SOMP and SSOMP ($\theta = 1.5$ and $\theta = 2$).

⁵ There is an interesting link with the work on the *Dantzig Detector* [2]. Under noisy conditions, this detector is used in sparse representation to first select the atoms whose representation coefficients are not zero, and then projects the signal orthogonally over their span, as we do here. This increases the performance of the algorithm.

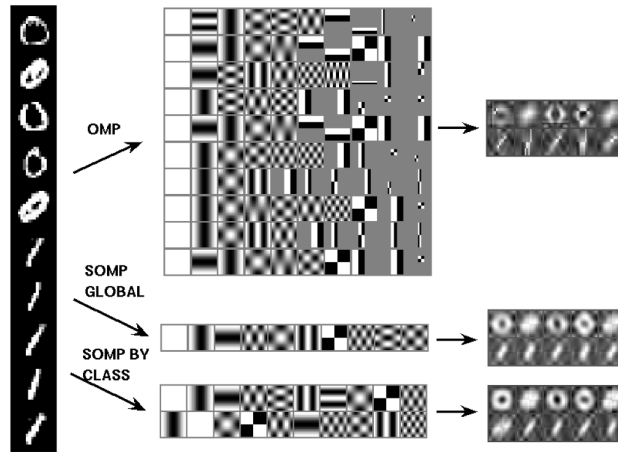


Fig. 4. *OMP vs SOMP vs Class-SOMP for a fixed dictionary. Left: test images. Top-center: atoms selected by OMP for the images of zeros and ones in the order they have been selected. Top-right: 10-sparse reconstructions from OMP coefficients. Center-center: atoms selected by SOMP for both of the classes together, in the order they have been selected. Center-right: reconstructions from 10-sparse SOMP coefficients. Bottom-center: atoms selected by SOMP for each one of the classes separately, in the order they have been selected. Bottom-right: reconstructions from 10-sparse Class-SOMP coefficients*

Each scenario is tested according to the following protocol: The testing ensemble is divided into ten blocks, each one containing the same number of images from each class. For each one of the blocks $\{b_i\}_{i=1}^{10}$, we train one ensemble of SVMs according to a one-against-one scheme with the coefficients of the rest of the blocks and using the parameters already fixed in advance (the training data is noiseless). We then repeat 5 times the following: Corrupt with noise or occlude the signals in the block b_i , and then project them orthogonally over the subset of the dictionary and perform classification over those coefficients (the testing data is now corrupted). We average the accuracies of the 5 repetitions and save them as accuracy of test over the block. We then average the accuracy results among the 10 executions. This procedure is a combination of 10-fold cross-validation with averaging due to the randomness of the noise and occlusion. The SVM parameters have been previously estimated through 10-fold cross-validation over a training set of 10000 images, different from the one used to perform SOMP or SSOMP. As we mentioned, the representation will have to deal with all the distortions introduced.

Both the results for noisy (Figure 5-left) and occluded (Figure 5-right) images show three major points. First of all, results are almost equivalent for all algorithms for high dimension (30 and 50). In those cases, most of the information has already been captured and thus improvement is not really possible. For complex datasets with less visual coherence, such high dimensions will capture too much intra-variance. For a dictionary of dimension 30, the pre-defined dic-

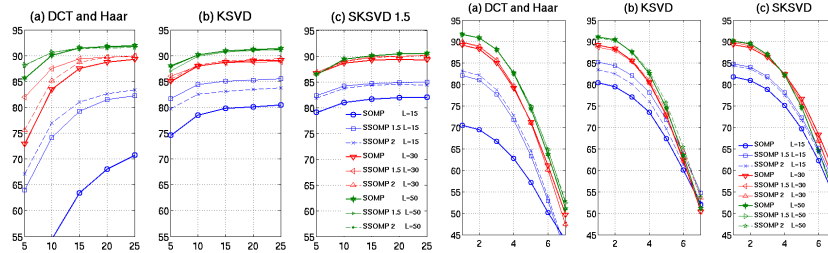


Fig. 5. Left: Classification under noisy conditions. SNR in dB in abscises (25 dB is noiseless), accuracy in ordinates. From left-to-right: (a) DCT and Haar, (b) KSVD and (c) SKSVD. In each one of the cases we perform analysis at $L = 15, 30, 50$. The selection is done by SOMP, SSOMP with $\theta = 1.5$, and SSOMP with $\theta = 2$. **Right:** Classification under occlusion conditions. Size of the occlusion in abscises, accuracy in ordinates. Same order as for noisy conditions. (This is a color figure.)

tionary shows a breakdown for $\text{SNR} \leq 10$, proving that learning a dictionary via SKSVD is more robust. When going from dimension 256 (the image dimension) to 15 (the dictionary dimension), SKSVD derived dictionaries provide significantly more robustness than the fixed one. For example, when decreasing the SNR from noiseless to 5 dB, the accuracy variation is smaller than 5% for SKSVD, whereas for the fixed dictionary is more than 15%. This implies that learning a dictionary provides a much more accurate description of the internal structure of the class, capturing much of the manifold the signals belong to.

Secondly, we verify the relevance of the discrimination power in the SSOMP. For dimension 15 for example, the SSOMP over the fixed dictionary achieves similar performance (slightly lower) to that of SKSVD in nearly noiseless situations. However, when noise is introduced, the representation remains highly un-robust, the performance falling by more than 10% relatively to the others for the same distortion level. It is clear that robustness comes from adaptation of the dictionary to the signals, learning is essential.

Thirdly, SOMP over SKSVD produces more accurate classification than over KSVD, since the discrimination component has been included in the dictionary learning itself. When coupled with SSOMP, both learning strategies yield similar results. This implies that the incorporation of the discrimination measure is highly important, but the gain in each of the stages seems to be limited.

We have also compared between learning the SKSVD dictionary and then performing dimensionality reduction and learning it directly at the proposed dimension. The obtained results (here omitted due to space limitations) show that not only incorporating the per-class SOMP is critical, but also that learning the set directly into the final dimension yields better results (with improvements of about 2%).

These results have clearly shown that learning dictionaries under combined discriminative and reconstructive constraints, as provided by our proposed SKSVD framework, is critical for signal classification, in particular for robustness under common distortions. Let us close with an illustration of the dictionaries learned by our proposed technique. Figure 6 presents those obtained for the experiments in Figure 4.

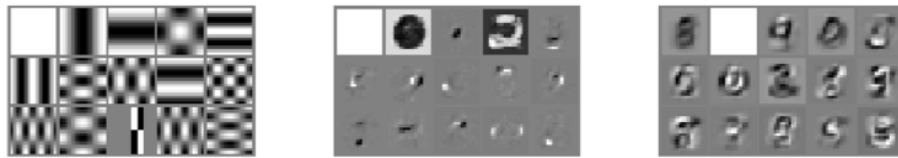


Fig. 6. Dictionaries of size 15 obtained through SSOMP $\theta = 2$ over DCT and Haar (left), KSVD (center) and SKSVD (right)

3.1 Natural Images

3.2 Working with Patches

Following the rich literature on object recognition and scene classification, we need to extend the present framework to work on local patches. For the digits dataset, we have performed that, and obtained the same robust classification results described in previous section, actually improving by 5% the classification (with a ground polynomial metric). In particular, all 8×8 patches are considered, and their signatures according to our proposed framework are clustered. Following this, digits are compared using a normalized sum-of-kernels distance between the corresponding vector signatures, where each coordinate in the vector is provided by the cluster center and cluster cardinality. This opens the door to exploit the proposed framework, for example, in the form of *bags-of-words* models, replacing standard SIFT-type of features by the ones explicitly learned for classification with our method.

Let us now present preliminary results in this direction for three classes shown in Figure 7 (top-left) from the Caltech Categories dataset. We first perform a standard key-point detector based on the Harris-Laplace approach. On these patches, we run our proposed learning technique, see Figure 7 (top-right) for examples of the learned dictionaries. The coefficients corresponding to the sparse representation over these learned dictionaries become the local discriminative feature descriptors for each patch (in contrast for example with SIFT). Once the local descriptors have been extracted, we perform standard K-means and obtain signatures for 40 clusters. The metric between those signatures is established with the Gaussian extension of the Earth Movers Distance (EMD), where the spread parameter is fixed as the average of all the EMDs, and the distance is the Euclidean one (linear kernel). We normalize the weights so that they have total weight one, thereby representing a probability distribution, where EMD can be interpreted in terms of the work to transform one probability distribution into the other. Finally we perform learning and classification with SVMs. For 100 samples from each class, the 10-fold cross validation accuracy has been of 94%. For 5 classes, adding leaves and cars rear, we obtained a classification of 89.7%. Following the success of this simple local study, which does not depend on ad-hoc features but formally computes the local descriptors following the energies provided by our framework, global structures (as currently done in the literature for the ad-hoc features), will prove to be further discriminant, inheriting also the robustness of the learned local representations. Further analysis will be carried out in this direction and results will be reported elsewhere.



Fig. 7. Examples from the classes faces, airplanes, and motorbikes from Caltech Categories (top-left); atoms learned with our technique from key patches (top-right).

4 Concluding Remarks

An energy based framework for learning dictionaries for simultaneous sparse signal representation and robust class classification has been introduced in this paper. This energy is minimized by a class-dependent simultaneous orthogonal matching pursuit interleaved with an efficient dictionary update. We have contributed to the understanding of learning sparse representations for signal classification, and showed the relevance of learning dictionaries to achieve accurate and robust classification. We demonstrated that performing simultaneous decomposition per class is essential in order to extract the internal structure of the class. The orthogonal projection over dictionaries increases robustness and unifies the representation of signals. We further demonstrated that learned dictionaries outperform fixed ones in classification tasks, in particular for distorted data and compact representations. Current work is concentrated in the topic of local patches, following the promising preliminary results described above.

Acknowledgments

We thank Julien Mairal for interesting discussion, inspiration, and for providing the initial KSVD code that was exploited in this paper. The work here reported is partially supported by NSF, ONR, ARO, DARPA, and NGA.

References

1. M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. SP*, 54:4311–21, 2006.
2. E. J. Candes and T. Tao. The dantzig selector: statistical estimation when p is much larger than n . Technical report, Caltech, 2005.
3. C. C. Chang and C. J. Lin. *LIBSVM*, 2001. <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
4. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.
5. A. Cohen, W. Dahmen, and R. DeVore. Compressed sensing and best k -term approximation. *preprint* <http://www.igpm.rwth-aachen.de/reports/pdf/IGPM260.pdf>, 2006.
6. M. Elad and M. Aharon. Image denoising via sparse and redundant representation over learned dictionaries. *IEEE Trans. Img Proc.*, 15(12):3736–44, December 2006.
7. M. Fritz, B. Leibe, B. Caputo, and B. Schiele. Integrating representative and discriminant models for object category detection.
8. A. Gilbert and J. Tropp. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans. Information Theory*, to appear.

9. R. Grosse, R. Raina, H. Kwong, and A. Y. Ng. Shift-invariant sparse coding for audio classification. In *Proc. Conf. on Uncertainty in AI*, 2007.
10. C.-W. Hsu and C.-J. Lin. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13(2):415–25, March 2002.
11. K. Huang and S. Aviyente. Sparse representation for signal classification. In *NIPS 19*, pages 609–616, 2007.
12. N. Jovic, B. Frey, and A. Kannan. Epitomic analysis of appearance and shape. In *Proc. ICCV*, 2003.
13. R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1137–1145, 1995.
14. E. Kokopoulou and P. Frossard. Supervised dimensionality reduction for joint compression and classification. <http://infoscience.epfl.ch/search.py?recid=91054>, 2007.
15. J. Lasserre, A. Kannan, and J. Winn. Hybrid learning of large jigsaws. In *Proc. IEEE CVPR*, 2007.
16. S. Lazebnik and M. Raginsky. Learning nearest-neighbor quantizers from labeled data by information loss minimization. In *Int. Conf. on AI and Stat.*, 2007.
17. B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. *Int. J. of Computer Vision*, 2007 (in press).
18. D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. of Computer Vision*, 60(2):91–110, 2004.
19. J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Learning discriminative dictionaries for local image analysis. In *Proc. CVPR*, Alaska, June 2008.
20. J. Mairal, M. Elad, and G. Sapiro. Sparse representation for color image restoration. *IEEE Trans. Img. Proc.*, January 2008.
21. S. Mallat and Z. Zhong. Matching pursuit in a time-frequency dictionary. *IEEE Trans. Signal Processing*, 41:617–43, 1992.
22. K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Trans. PAMI*, 27(10):1615–1630, 2005.
23. B. Olshausen, P. Sallee, and M. Lewicki. Learning sparse multiscale image representations. *NIPS*, 15:1327–1334, 2003.
24. Y. Pati, R. Rezaifar, and P. Krishnaprasad. Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. In *Proc. Assilomar Conference on Signals, Systems and Computers*, November 1993.
25. G. Peyre. Sparse modeling of textures. In *Preprint Ceremade 2007-15*, www.ceremade.dauphine.fr/~peyre/publications/.
26. M. Ranzato, F. Huang, Y. Boureau, and Y. LeCun. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *CVPR*, 2007.
27. M. Ranzato, C. Poultney, S. Chopra, and Y. LeCun. Efficient learning of sparse representations with an energy-based model. In *NIPS*, pages 1137–1144, 2007.
28. S. Roth and M. J. Black. Fields of experts: A framework for learning image priors. In *CVPR 2005*, volume 2, pages 860–867 vol. 2, San Diego, USA, June 2005.
29. J. A. Tropp. Algorithms for simultaneous sparse approximation. part II: Convex relaxation. *Signal Processing*, 86(3):589–02, 2006.
30. J. A. Tropp, A. C. Gilbert, and M. J. Strauss. Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit. *Signal Processing*, 86(3):572–88, 2006.
31. S. A. J. Winder and M. Brown. Learning local image descriptors. In *CVPR 2007*, pages 1–8, June 2007.
32. Y. Ma et al. <http://perception.csl.uiuc.edu/recognition/home.html>. 2007.