A METHOD FOR DISCRIMINATIVE DICTIONARY LEARNING WITH APPLICATION TO PATTERN RECOGNITION

Román E. Rolón[♭], Leandro E. Di Persia[♭], Hugo L. Rufiner[♭, †] and Rubén D. Spies[‡]

[♭] *Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional (sinc(i), UNL-CONICET), Facultad de Ingeniería y Ciencias Hídricas, Univ. Nacional del Litoral, Santa Fe, Argentina, rrolon@sinc.unl.edu.ar, ldipersia@sinc.unl.edu.ar, lrufiner@sinc.unl.edu.ar*

[†] *Laboratorio de Cibernética, Facultad de Ingeniería, Univ. Nacional de Entre Ríos, Argentina.*

[‡] *Instituto de Matemática Aplicada del Litoral (IMAL), UNL-CONICET, Santa Fe, Argentina, rspies@santafe-conicet.gov.ar*

Abstract: Pattern recognition is a scientific discipline whose purpose is the classification of objects into different categories or classes. Object categorization deals with the detection or recognition of “generic” categories, reason for which it known as “generic object recognition”. In this article, sparse representation of signals in terms of a discriminative multi-class dictionary for image recognition is presented. A sparse representation approximates an input signal over a linear combination of a few atoms of the given dictionary. A balanced set of input signals selected from the Caltech 101 database is used for learning the discriminative dictionary. The sparse vectors are then used as input of a multi-class classifier. The proposed method shows improvements over the standard KSVD method.

Keywords: *Dictionary learning, Inverse problems, Discriminative information*
2000 AMS Subject Classification: 21A54 - 55P54

1 INTRODUCTION

The problem of sparse representation of signals consists of obtaining representations of such signals by means of a linear combination of only a few atoms of an appropriately constructed dictionary [1]. Some of the advantages of sparse representations are super resolution, robustness to noise and dimension reduction, among others. The sparse representation problem can be divided into two separate sub-problems: an inference and a learning problem. The first one, which is usually called “sparse coding”, consists of selecting a set of representation vectors $\{\mathbf{a}_i\}$ satisfying a given sparsity constraint. The second problem, which is quite often more complex, consists of finding an “optimal” dictionary Φ for representing a given set of signals $\{\mathbf{x}_i\}$. The formulation of the learning problem focuses only on minimizing the reconstruction error without taking into account the discriminative classification power of the dictionary [1]. For that reason, some authors have proposed different supervised approaches in order to take advantage of the discriminative power of the dictionary [2], [3]. In such supervised approaches the dictionary and a linear classifier are simultaneously optimized.

In a previous work [4], two different methods for identifying the most discriminative atoms in an overcomplete dictionary were presented. A method for learning discriminative dictionaries to be used for classification tasks is presented in this work.

The article is organized as follows: the method used for learning discriminative dictionaries is described in Section 2. Experiments and results are presented and discussed in Section 3. Finally, conclusions are presented in Section 4.

2 METHODS

2.1 SPARSE CODING

A sparse representation of a given data matrix of N input signals $X \doteq [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N]$ where $\mathbf{x}_i \in \mathbb{R}^n$ in terms of a given dictionary $\Phi \in \mathbb{R}^{N \times M}$ (with $M \geq N$), whose columns ϕ_j are sometimes called atoms, can be obtained by minimizing the following problem,

$$A^* = \underset{A}{\operatorname{argmin}} \sum_{i=1}^{N} \left(||\mathbf{x}_i - \Phi \mathbf{a}_i||_2^2 + \lambda ||\mathbf{a}_i||_1 \right), \tag{1}$$

where λ is a sparsity constraint factor and the terms $||\mathbf{x}_i - \Phi \mathbf{a}_i||_2$ are the reconstruction errors. Finally, each input signal $\mathbf{x}_i$ is approximated by a linear combination of only a few atoms of the dictionary in an appropriate way.

2.2 Dictionary learning

The aim of dictionary learning is to obtain an efficient dictionary that provides a good representation for most of the signals under study. The dictionary learning problem can be stated as follows:

$$< \Phi^*, A^* >= \underset{\Phi, A}{\text{argmin}} \sum_{i=1}^{N} \left(||\mathbf{x}_i - \Phi \mathbf{a}_i||_2^2 + \lambda ||\mathbf{a}_i||_1 \right). \tag{2}$$

Problem (2) is convex in each one of the variables Φ and A, individually, but not simultaneously convex [5]. Therefore the dictionary learning problem is usually iteratively and sequentially solved, by first optimizing in Φ (while holding A fixed) and then optimizing in A (while holding Φ fixed), so proceeding until a prescribed stop criteria is met. Clearly, formulation (2) minimizes the reconstruction error and this formulation does not take into account the discriminative information of the dictionary, which is completely neglected for the classification task.

2.3 Discriminative dictionary learning

In a previous work [4] we have introduced a novel approach for identifying the most discriminative atoms in a binary classification problem. In this work a method for learning a discriminative dictionary to be used for solving a multi-class problem is proposed. The activation frequency of the atoms is denoted by η_κ^j where j represents the activation of atom ϕ_j for input signals labeled as belonging to class κ. With η_κ^j we shall denote the number of times that the atom ϕ_j is used to represent data belonging to class κ. Suppose that the atom ϕ_j has a very high activation for the class "κ" but very low activation for the remaining classes. In such a case the atom ϕ_j is considered to be highly discriminative for classifying elements belonging to the class "κ". Otherwise, if the atom ϕ_j has similar activation for the class "κ" and for the remaining classes, then the atom is considered as not carrying any significant discriminative information.

The discriminative approach begins by defining a matrix $D \in \mathbb{R}^{M \times \kappa}$, which represents a measure of the discriminative power of the atoms:

$$D(j,k) = \frac{1}{N} \left(\eta_k^j N_{\kappa \neq k} - \sum_{\kappa \neq k} \eta_k^j N_k \right), \tag{3}$$

where N_k denotes the number of input signals belonging to the class "k" while $N_{\kappa \neq k} (= n - N_k)$ denotes the number of signals not belonging to that class. The elements in the rows of D represent the difference of activations of the atoms in the dictionary and the elements of its columns represent the difference of activation frequencies (see Eq. (3)). Clearly $D(j,k)$ will be positive and large if the j^{th}-atom is much more active in one class than in the others. Otherwise, if the j^{th}-atom has similar activation frequencies for all the classes, then $D(j,k)$ will be small or even negative.

We describe now the building steps of the discriminative dictionary learning method together with the corresponding lines of its implementation algorithm (Algorithm 1). Let X be the training data, p_0 the sparsity level, I the final number of discriminative atoms for each class and κ the number of classes. The algorithm starts by learning a dictionary Φ by using the unsupervised KSVD algorithm [1] (line 4). Then each representation matrix A_κ is obtained by applying a greedy pursuit algorithm called OMP [6] (line 5). Then, matrix D is obtained according to Ec. (3) (line 6). An atom selection process follows by selecting the most discriminative ones for each one of the classes. Those discriminative atoms constitute the columns of the matrix Φ_d. It could happen that no discriminative atoms are available for certain classes. In such a case, the corresponding columns of Φ_d are represented by vectors of zeros and the indices of those classes are saved into the vector Rem (line 7). Finally a new data matrix $\hat{X}$ is constructed by removing all the input signals of X belonging to the classes corresponding to the discriminative atoms (line 12). This process is repeated until all of the discriminative atoms are acquired.

The idea behind this approach is to learn a discriminative dictionary where the activations of the atoms contain significant information to be used for classification. Thus, a sparse version of a multi-class linear discriminant analysis (LDA) has been chosen for classification [7].

Algorithm 1 DKSVD

```
procedure DKSVD($X, p_0, I, \kappa$)
    $inc = 1$
    for $i \leftarrow 1, I$ do
        $\Phi \leftarrow$ KSVD($X, p_0$)
        Get sparse matrix $A = [A_1\ A_2\ A_3\ \cdots\ A_\kappa]$ that accomplish $X = \Phi A$
        Get $D$ according to Ec. (3)
        Get $\Phi_d$ and $Rem$
        if $Rem = \emptyset$ then
            $inc = 0$
        end if
        while $inc = 1$ do
            Get $\hat{X}$ by removing the input signals corresponding to the discriminative atoms
            $\Phi \leftarrow$ KSVD($\hat{X}, p_0$)
            Get sparse matrix $A$ that accomplish $\hat{X} = \Phi A$
            Get $D$ according to Ec. (3)
            Get $\Phi_d$ and $Rem$
            if $Rem = \emptyset$ then
                $inc = 0$
            end if
        end while
        $\Phi_D \leftarrow [\Phi_D\ \Phi_d]$
        Remove randomly a signal per class
    end for
    $\Phi_D \leftarrow [\Phi_{c1}\ \Phi_{c2}\ \cdots\ \Phi_{c\kappa}]$ where $\Phi_{c\kappa} = [\phi_{c1}^{1}\ \phi_{c2}^{2}\ \cdots\ \phi_{c\kappa}^{I}]$
    return $\Phi_D$
end procedure
```

3 Experiments and results

The Caltech 101 database is used for this work [8]. This database is widely used and it contains 9144 images corresponding to 101 different objects (classes) with an extra background class. The images have an average size of 60,000 pixels (300 × 200). The number of images belonging to each class varies from 31 (inline_skate) to 800 (airplanes). The images corresponding to the background and faces classes were not taken into account for this work. Thus, a total of 100 classes were considered.

To balance the database, a total of 30 images from each class were randomly selected, of which 15 were used for training and the remaining ones for testing purposes. In the pre-processing stage, the region of interest of each image was automatically cropped, converted into gray scale, normalized using histogram normalization, resized to 32×32 and finally converted into vectors of length 144 using Principal Components Analysis (PCA) [9].

The matrix of input signals $X \in \mathbb{R}^{144\times3000}$ was built by staking side-by-side the input matrices corresponding to each class, i.e. $X = [X_{c1}, X_{c2}, \cdots, X_{c100}]$. Next, the signal matrices $X_{train} \in \mathbb{R}^{144\times1500}$ and $X_{test} \in \mathbb{R}^{144\times1500}$ were constructed by randomly selecting signals from each class in X.

The first step of our method is to obtain the initial dictionary. In this step no information about classes is required. To accomplish this task the standard KSVD algorithm is used [1]. In the following step, the representation coefficients of X_{train} are obtained by applying the OMP algorithm [6]. By considering the representation coefficients, the matrix D which contains significant information about the discriminative power of each atom in the initial dictionary, is constructed. The image appearing on the left of Figure 1 shows a representation of the matrix D. It can be seen that, for most of the classes, the atoms have a high variability of activation frequency. On the right of Figure 1, a 3-D representation of matrix D, where the elements of each column are now shown in decreasing order of magnitude, is presented. It can be clearly seen that, for instance, the 258^{th}-atom contains high discriminative information for class 3

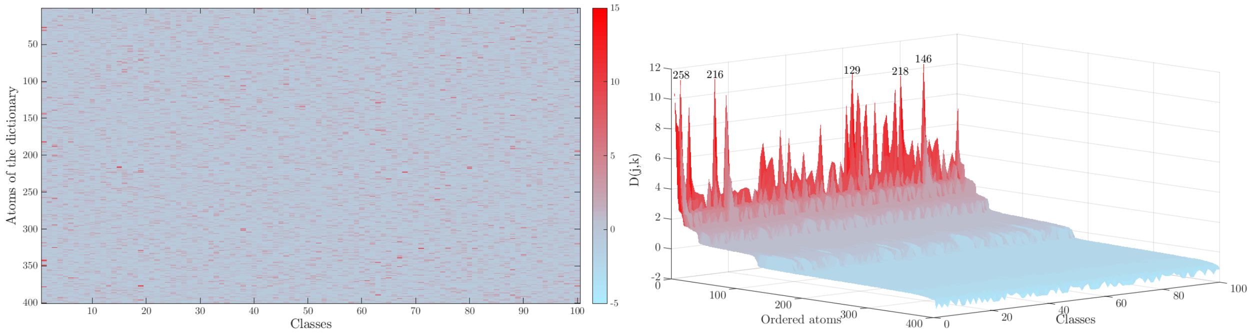

Figure 1: Difference of activation frequency. Matrix D (left); Matrix D with decreasingly reordered columns (right).

($D(258, 3) = 11.48$). On the studied database, by considering 15 input signals for training and testing, a classification accuracy of 24,6% was obtained. Compared with the use of a dictionary learned by means of the standard KSVD, whose classification accuracy resulted in 21,3%, our result is encouraging.

4 Conclusions

A new approach to pattern recognition using sparse representations was presented. The results show that the design of a discriminative dictionary is a suitable technique to be used for multi-class classification tasks. Although far more research is needed in order to improve the classification performances, the approach constitutes a promising method for learning a discriminative dictionary. For future work we propose to use over-sampling techniques [10] for increasing the size of the database. The effects of using different types of classifiers will also be studied.

Acknowledgments

This work was supported in part by Consejo Nacional de Investigaciones Científicas y Técnicas, CONICET, through PIP 2014-2016 Nro. 11220130100216-CO and PIP 2012-2014 Nro. 114 20110100284-KA4, by the Air Force Office of Scientific Research, AFOSR / SOARD, through Grant FA9550-14-1-0130 and by Universidad Nacional del Litoral through project CAI+D 501 201101 00519 LI.

References

[1] Aharon M., Elad M. and Bruckstein A. *K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation*. IEEE Transactions on Signal Processing, Vol. 54, (2006), pp. 4311-4322.

[2] Pham D. and Venkatesh, S. *Joint learning and dictionary construction for pattern recognition*. In proceedings CVPR Workshop on Generative-Model Based Vision, (2008).

[3] Jiang Z., Lin Z. and Davis L. *Learning a discriminative dictionary for sparse coding via label consistent k-svd*. In proceedings CVPR Workshop on Generative-Model Based Vision, (2011).

[4] Rolón R., Larrateguy L., Di Persia L., Spies R. and Rufiner L. *Discriminative methods based on sparse representations of pulse oximetry signals for sleep apnea-hypopnea detection*. Biomedical Signal Processing and Control, Vol. 33, (2017), pp. 358-367.

[5] Guo H., Jiang Z. and Davis L. *Discriminative Dictionary Learning with Pairwise Constraints*. In proceedings ACCV Asian Conference on Computer Vision, Vol. 7724, (2012), pp. 328-342.

[6] Pati Y.C., Rezaiifar R. and Krishnaprasad P.S. *Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition*. IEEE Asilomar Conference on Signals, Systems and Computers, Vol. 1, (1993), pp. 40-44.

[7] Clemmensen L., Hastie T., Witten D. and Ersbøll B. *Sparse Discriminant Analysis*. Technometrics, 53, (2011), pp. 406-413.

[8] Fei-Fei L., Fergus R. and Perona P. *Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories*. In proceedings CVPR Workshop on Generative-Model Based Vision, (2004).

[9] Jolliffe I.T. *Principal Component Analysis*. Springer, 2nd ed, (2002).

[10] Chawla N., Bowyer K., Hall L. and Kegelmeyer P. *SMOTE: Synthetic Minority Over-sampling Technique*. Journal of Artificial Intelligence Research, Vol. 16, (2002), pp. 321-357.

[11] Crammer K. and Singer Y. *On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines*. Journal of Machine Learning Research 2, (2001), pp. 265-292.