DISCRIMINATIVE SPARSE REPRESENTATIONS FOR IMAGE CLASSIFICATION

By

Suhaily Cardona-Romero

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

MASTER OF SCIENCE

Electrical Engineering

2012

ABSTRACT

DISCRIMINATIVE SPARSE REPRESENTATIONS FOR IMAGE CLASSIFICATION

By

Suhaily Cardona-Romero

Sparse representations and compressed sensing (CS) are two methods that have drawn the attention of the signal processing community due to their ability to reduce the dimensionality of signals while preserving enough information for signal representation. However, these compact representations do not necessarily preserve the most discriminative aspects of the signal. This thesis addresses this issue by developing a new discriminative framework to obtain a compact representation with high discriminative information for image classification applications.

The first part of this thesis presents a greedy algorithm inspired by CoSaMP with the inclusion of a new cost function that quantifies the tradeoff between discrimination power and sparsity. The inclusion of this cost function helps to select a small number of atoms from an overcomplete dictionary that produces discriminative sparse representations of images from different classes. Through experiments, it was shown that such representations can be used as features to classify new sample images even under noisy environments or missing pixels.

The second part of this thesis proposes a method to obtain discriminative measurements from CS and is motivated by the fact that the presence of irrelevant features may reduce the classification accuracy. To address this issue, a feature selection step was added to CS to eliminate irrelevant features from the measurements. As a result of the elimination of such features, an improvement in the classification accuracy is observed. In conclusion, it was demonstrated that a subset of incoherent projections with high discrimination power performs better than the whole set of CS measurements for classification purposes.

To my parents, three brothers and boyfriend
for their unconditional love and support

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Selin Aviyente, for her guidance and support during my graduate studies. Not only did she guide me through my thesis but she was an integral part in providing me with the knowledge that I needed to be able to do my work. Also, I want to thank Dr. Lalita Udpa for her time and support while being a part of my committee. I would also like to take this opportunity to thank my lab mates: Marcos Bolanos, Ying Li and Ali Mutlu, for helping me when I needed it.

In addition, I am very grateful to Dr. Barbara O'Kelly, Dr. Percy Pierre and the Sloan Engineering Program for the opportunity they gave me to pursue my graduate studies at MSU. Through their seminars and meetings not only did I get the opportunity to network with other people but I was also able to learn some valuable information for my life as a graduate student. Also, I want to thank them for their support and funding provided during these two years.

Finally, I would like to thank my parents, Jose and Migdalia, and my three brothers, Jose, Eduar and Anthony, for their unconditional love and support through my entire life. I want to thank my parents for all the effort they made to give me the formation that I needed in order to succeed in life. Also, I want to thank my beloved boyfriend, Eduardo E. Montalvo, for all his support, encouragement and love. Thanks to his company and sense of humor, this chapter of my life was more pleasant. When I was going through hard times, he motivated me and helped me to continue achieving my goals. Thank you all, without your help this would have not been possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

BiCGStab	Biconjugate Gradient Stabilized
BP	Basis Pursuit
COIL	Columbia Object Image Library
CoSaMP	Compressive Sampling Matching Pursuit
DiscCoSaMP	Discriminative CoSaMP
DWT	Discrete Wavelet Transform
i.i.d.	Independent Identically Disctributed
KPCA	Kernel Principal Component Analysis
LASSO	Least Absolute Shrinkage and Selection Operator
LDA	Linear Discriminative Analysis
LLE	Locally-Linear Embedding
MP	Matching Pursuit
OMP	Orthogonal Matching Pursuit
PC	Principal Component
PCA	Principal Component Analysis
RecCoSaMP	Reconstructive CoSaMP
RIP	Restricted Isometric Property
SNR	Signal-to-Noise Ratio
SVM	Support Vector Machine

CHAPTER 1

INTRODUCTION

Through the years, the innovations in sensor technology have led to the collection of massive amounts of data with high dimensionality. The analysis of large amounts of high dimensional data can pose a problem in the computation time for many applications in the area of image processing such as texture classification, object detection, and recognition. This problem can be addressed by eliminating dimensions that seem to be redundant or irrelevant to the desired application. Dimensionality reduction is often used as a preprocessing technique that looks for a low dimensional representation from a high dimensional signal, using linear or nonlinear methods, such that the structure of the signal is preserved. The goal of this thesis is to address the dimensionality reduction problem to extract a low dimensional feature vector with high discrimination power for image classification applications.

The most widely used linear methods for dimensionality reduction are Principal Component Analysis (PCA) and Linear Discriminative Analysis (LDA). The goal of PCA is to, citing Jolliffe [1]:

> "…reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. This is achieved by transforming to a new set of variables, the principal components (PCs), which are uncorrelated, and which are ordered so that the first *few* retain most of the variation present in *all* of the original variables."

The implementation of PCA results in orthogonal projections that represent the data set with a few components or dimensions, i.e. PCs, without taking into account the differences between the

classes making such a projection unsuitable for discrimination purposes. Contrary to PCA, LDA reduces the dimensionality by seeking the best projection that separates the different classes in the data set by maximizing the between-class scatter while at the same time minimizing the within-class scatter. However, LDA is known to be very sensitive to noise in the data.

Nonlinear methods used to reduce the dimensionality of a data set include isomaps, locally-linear embedding (LLE) and kernel mapping [2, 3]. Isomaps reduce the dimensionality by mapping high dimensional data into a lower dimensional space while it preserves the neighborhood distances and the geodesic distances between all pair of features. Similar to isomaps, LLE reduces the dimensionality by preserving the neighborhood distances; the difference is that LLE tries to minimize the least squares error of the geodesic distances [2]. Kernel methods map the data from the high dimensional space $\Re^d$ into a dot product space F via a nonlinear function, using methods such as Kernel Principal Component Analysis (KPCA) [3] which is a generalization of PCA.

In recent years, sparse representations and compressed sensing have been used to reduce the dimensionality of data sets. Sparse representations reduce the dimensionality by projecting the data against a small number of elements (called atoms) from an overcomplete dictionary such that the reconstruction error is minimal. In this case, the atoms to obtain the low dimensional projection of the data can be selected through the implementation of greedy algorithms such as Matching Pursuit [30], Orthogonal Matching Pursuit [31], Compressive Sampling Matching Pursuit [32] and convex optimization methods such as the l_1-minimization problem and LASSO [27, 34]. On the other hand, compressed sensing represents a sparse signal with a small number of measurements through random projections. Using this method, the number of measurements obtained is much less than the number of samples required from the Nyquist sampling theorem

while still allowing the reconstruction of the original signal with no or little loss of information. These are two methods that have been widely used for signal representation and compression purposes.

Recently, the signal processing community has been interested in expanding these methods to applications such as object detection, face recognition and image classification. In [4], a method based on sparse representation for text detection is proposed using a learned dictionary obtained using text-like edge features extracted from a database of text images. This dictionary in conjunction with OMP is used to obtain sparse representations of the edge features extracted from the test images using a 16×16 window. The sparse representation obtained from the 16×16 window is detected as text using a threshold based on the number of nonzero elements in the representation. A similar method for text detection can be found in [5], where two different dictionaries are learned (one from the text images and the other from the background) and the subimage within the 16×16 window is projected using each dictionary resulting into two different projections. The subimage is detected as text or background depending on which projection produces the smaller reconstruction error. Sparse representations have also been used for the detection of humans in images [6] thanks to its multi-scale nature. In [32], Wright et al. proposed an algorithm for face recognition that use the training images as the atoms of the dictionary to solve the sparse representation problem and classify the test image to the class of the training image (atom) that minimizes the reconstruction error. In [7], it is shown that compressed sensing measurements can be used to subtract the background of test images. Given a background scene image x_b and a test image x_t of the same scene, any discrepancy present can be obtained by the pixel-wise subtraction of both images ($x_b - x_t$). However, if the images are available in the compressed space, Cevher et al. showed that the discrepancies can be

obtained by reconstructing the difference between the background measurements y_b and the test measurements y_t $(y_b - y_t = \Psi x_b - \Psi x_t)$.

In this thesis, a new theoretical framework is developed for using sparse representations and compressed sensing for image classification applications. Currently, sparse representations are mostly limited to reconstructive representations of signal. For classification purposes, it is more important to obtain discriminative sparse representations instead of reconstructive ones, i.e. minimizing reconstruction error. To achieve this goal a new optimization cost function that combines discrimination power and sparsity is proposed along with a modified greedy pursuit algorithm. This cost function allows a tradeoff between discrimination and sparsity that helps with the selection of the smallest number of atoms from an overcomplete dictionary that best discriminate a set of training images. The indices of the atoms identified as the most discriminative in the training stage are used to obtain the discriminative sparse representation of the test images. The performance and robustness of the proposed method is evaluated by performing classification experiments with two different image databases under different levels of Gaussian noise and different sizes of occlusion. The first experiment is performed with a standard image database with low intra-class variability and objects in a black background. The results of this experiment are compared to a modified OMP algorithm. The proposed method selects multiple atoms per iteration as opposed to OMP which selects a single atom at each iteration. This modification enables the proposed algorithm to select features in a more computationally efficient way while at the same time achieving comparable or better classification accuracy than the modified OMP algorithm. The second experiment uses a more challenging database with high intra-class variability to show the effect of class variability on sparseness. In this experiment, a comparison with LDA is presented to illustrate the superior

performance of the proposed algorithm in terms of accuracy and sparsity over conventional methods.

Similar to the sparse representation methods, compressed sensing has been mostly used for signal compression and reconstruction. Recently, Haupt et al. have shown that these measurements can also be used for signal classification purposes [63]. In addition, they presented and validated a theoretical misclassification bound that shows that the error probability decays exponentially as the number of measurements increase. In this thesis, this approach is extended to image classification by using a cost function based on the Fisher score to select the most relevant compressed sensing measurements (features) for classification purposes instead of using all the measurements. The cost function proposed will measure how well the compressed sensing measurements maximize the between-class scatter while at the same time minimize the within-class scatter. The motivation to include this cost function comes from the study made in [46] where Almaullin and Dietterich showed through experiments that the presence of redundant or irrelevant features can drop the classification accuracy significantly. With the inclusion of this cost function, it is expected to obtain higher classification accuracy using only a small number of measurements rather than using the whole set of measurements.

The organization of this thesis is as follows. Chapter 2 briefly reviews some basic concepts on wavelets, overcomplete dictionaries, sparse representations, compressed sensing and some common greedy algorithms used to solve the sparse representation problem as well as some applications in these areas. Chapter 3 presents a new greedy algorithm to classify images using sparse representations. A new cost function combining discrimination power and sparsity to obtain discriminative sparse representations is proposed. The performance and the robustness of the proposed algorithm are evaluated through experiments and it was shown that the algorithm

can work under noisy and occluded environments and can perform better than existing dimensionality reduction methods. In addition, the performance of the proposed algorithm is compared with LDA and a modified version of OMP. Chapter 4 presents a feature selection method from compressed sensing samples to improve the classification accuracy without the presence of irrelevant or redundant features.

CHAPTER 2

BACKGROUND

This chapter presents basic concepts and current applications in the area of transform based image feature extraction and classification that will serve as background for the work presented in this thesis. First, a brief overview of wavelets and wavelet transforms with some applications in the area of image processing will be presented. After this overview, the extension of wavelets to sparse representations using overcomplete dictionaries will be introduced. Then, some of the optimization methods and greedy algorithms that have been proposed to solve the sparse representation problem will be described. Finally, some background information about compressed sensing will be presented as well as some applications in the area of signal/image processing.

2.1 WAVELETS

An orthogonal family of wavelets is a collection of orthogonal basis functions created by dilating and translating a "mother wavelet" $\psi(t)$ [8]:

$$\psi_{j,n}(t) = \frac{1}{\sqrt{2^j}} \psi\left(\frac{t - 2^j n}{2^j} \right) \quad \text{for } (j,n) \in Z^2 \tag{2.1}$$

where j defines the scale of the basis and n defines its translation. In the same way, an orthogonal scaling family can be defined as:

$$\varphi_{j,n}(t) = \frac{1}{\sqrt{2^j}} \varphi\left(\frac{t - 2^j n}{2^j} \right) \quad \text{for } (j,n) \in Z^2 \tag{2.2}$$

where $\varphi(t)$ corresponds to the scaling function and $\varphi_{j,n}(t)$ is orthogonal to $\psi_{j,n}(t)$. These two functions usually are used together to perform multiresolution analysis of signals where the representations obtained with scaling functions $\varphi_{j,n}(t)$ correspond to the low frequency information in the signal and the representations obtained with wavelet functions $\psi_{j,n}(t)$ correspond to the high frequency information. Some of the most common orthogonal wavelet families include Haar, Daubechies, Symlets and Coiflets. Examples of 1D wavelets and scaling functions can be seen in Figure 1.

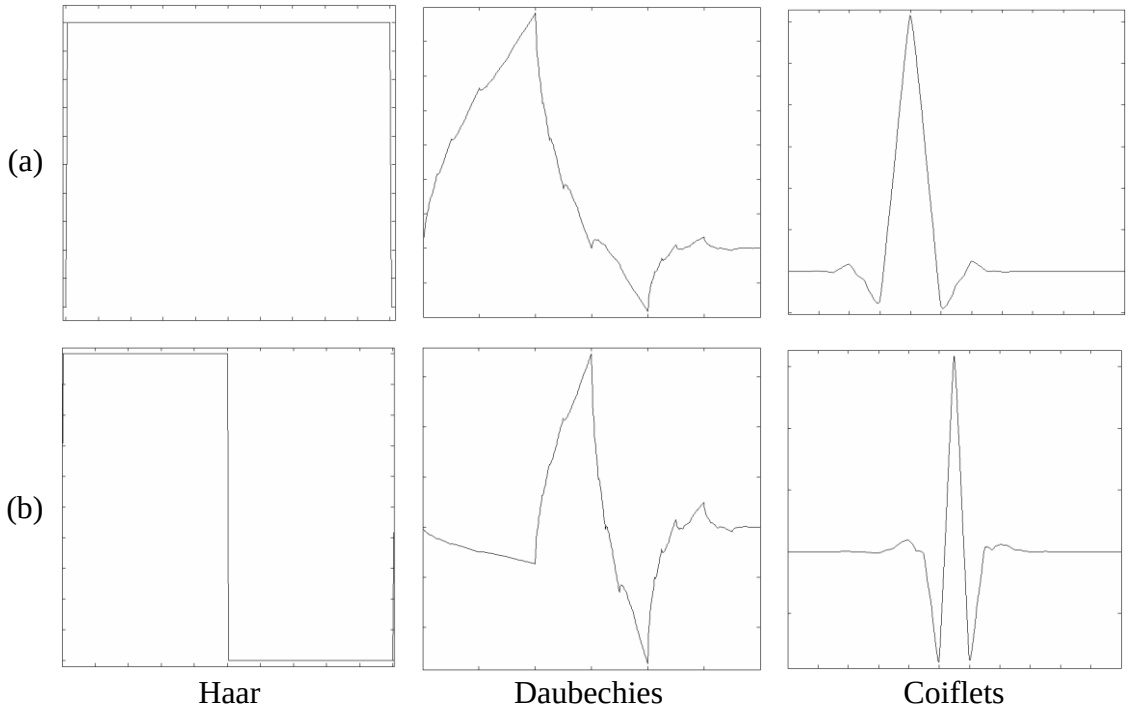

Figure 1. Scaling (a) and Wavelet (b) functions for Haar (left), Daubechies (middle) and Coiflets (right) wavelet families

A decomposition of any finite energy signal f can be obtained through the inner product of the signal and the scaling and wavelet families respectively as:

$$\boldsymbol{a}_j(n) = \left\langle \boldsymbol{f}, \varphi_{j,n} \right\rangle = \sum_{l=-\infty}^{\infty} \boldsymbol{f}(l)\varphi^*_{j,n}(l) \tag{2.3}$$

$$\boldsymbol{d}_j(n) = \left\langle \boldsymbol{f}, \psi_{j,n} \right\rangle = \sum_{l=-\infty}^{\infty} \boldsymbol{f}(l)\psi^*_{j,n}(l) \tag{2.4}$$

where $\boldsymbol{a}_j(n)$ corresponds to the approximation coefficients (low frequency information) and $\boldsymbol{d}_j(n)$ corresponds to the wavelet/detail coefficients (high frequency information). Therefore, an orthogonal expansion of $f(t)$ can be obtained as:

$$f(t) = \sum_n a_{j0}(n)\varphi_{j0,n}(t) + \sum_{j=j_0}^{\infty}\sum_k d_j(n)\psi_{j,n}(t) \tag{2.5}$$

Since the set $\{\varphi_{j_0,n}(t), n \in Z\}$ spans the same subspace as $\{\psi_{j,n}(t), j < j_0, n \in Z\}$ [9] this expansion can also be obtained as:

$$f(t) = \sum_j \sum_k d_j(n)\psi_{j,n}(t) \tag{2.6}$$

A fast implementation of the wavelet decomposition can be obtained by applying the filter bank theory. Filter banks decompose a signal into approximation and details coefficients by convolving the signal with low-pass filters $\boldsymbol{h}(n)$ and high-pass filters $\boldsymbol{g}(n)$ and downsampling the output by two as:

$$\boldsymbol{a}_1(n) = \sum_{l=-\infty}^{\infty} \boldsymbol{h}(l-2n)\boldsymbol{a}_0(l) \quad \text{and} \quad \boldsymbol{d}_1(n) = \sum_{l=-\infty}^{\infty} \boldsymbol{g}(l-2n)\boldsymbol{a}_o(l) \tag{2.7}$$

where $\boldsymbol{a}_0(l)$ corresponds to the input signal or the approximation coefficients at the highest scale. Multiple levels of decomposition can be achieved by iterating the analysis stage on the approximation coefficients as shown in Figure 2. This decomposition provides a description of

the signal at different scales such that coarse and fine features can be obtained simultaneously to analyze the signal.

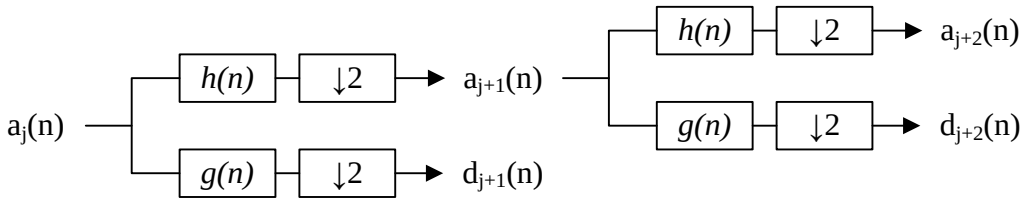

Figure 2. Analysis filter bank for 1D signals

This decomposition has been extended to two dimensional signals such as images using separable wavelets. The 1D scaling and wavelets functions are extended to 2D through the combination of products of these functions as follows:

$$\varphi(x_1, x_2) = \varphi(x_1)\varphi(x_2)$$

$$\psi^H(x_1, x_2) = \psi(x_1)\varphi(x_2)$$

$$\psi^V(x_1, x_2) = \varphi(x_1)\psi(x_2)$$

$$\psi^D(x_1, x_2) = \psi(x_1)\psi(x_2) \tag{2.8}$$

where ψ^H, ψ^V and ψ^D correspond to wavelets in the horizontal, vertical and diagonal directions, respectively. Given 2D separable scaling and wavelet functions, respectively defined as:

$$\varphi_{j,m,n}(x_1, x_2) = \frac{1}{2^j}\varphi\left(\frac{x_1 - 2^j m}{2^j}, \frac{x_2 - 2^j - n}{2^j}\right)$$

$$\psi^i_{j,m,n}(x_1, x_2) = \frac{1}{2^j}\psi^i\left(\frac{x_1 - 2^j m}{2^j}, \frac{x_2 - 2^j - n}{2^j}\right) \text{ for } i = \{H, V, D\} \tag{2.9}$$

An image $\boldsymbol{F}(x_1, x_2)$ of size $M \times N$ can be decomposed into the approximation and detail coefficients as:

$$\boldsymbol{a}(j_0, m, n) = \frac{1}{\sqrt{MN}} \sum_{x_1=0}^{M-1} \sum_{x_2=0}^{N-1} \boldsymbol{F}(x_1, x_2) \varphi_{j_0,m,n}(x_1, x_2)$$

$$\boldsymbol{d}^i(j, m, n) = \frac{1}{\sqrt{MN}} \sum_{x_1=0}^{M-1} \sum_{x_2=0}^{N-1} \boldsymbol{F}(x_1, x_2) \psi^i_{j,m,n}(x_1, x_2) \quad \text{for } i = \{H, V, D\} \quad (2.10)$$

respectively.

Similar to the 1D case, a filter bank can be implemented to obtain the approximation and details coefficients of an image by applying high-pass ($\boldsymbol{g}$) and low-pass ($\boldsymbol{h}$) filters to the rows and the columns of the input image and dowsampling the outputs by two as shown in Figure 3.

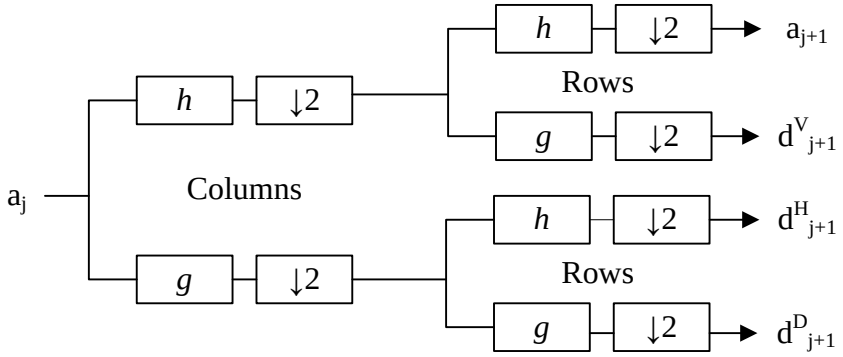

Figure 3. Analysis filter bank for 2D signals

In this case, the results are subimages with the approximation and detail coefficients of the original image. To obtain multiple levels of decomposition, the same process can be applied to the approximation coefficients which results in four additional subimages (Figure 4). For N levels of decomposition, the image is decomposed into $3N + 1$ subimages.

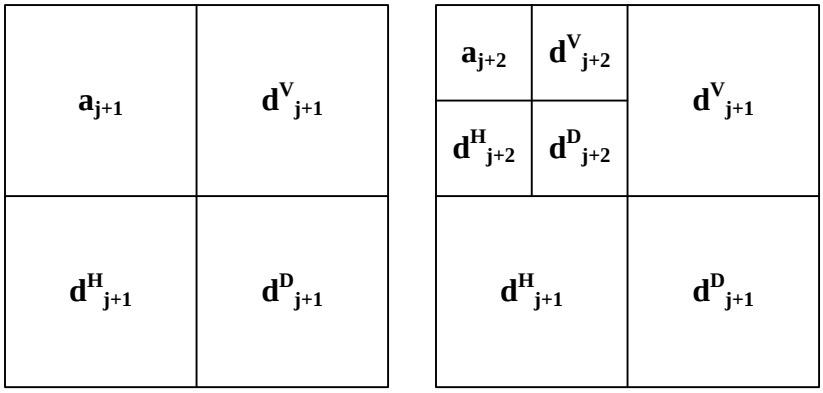

Figure 4. Subimages for one (left) and two (right) levels of decomposition

Wavelet decomposition has been very popular in areas such as signal processing, computer vision, pattern recognition and medical applications, among others. This popularity is due to its ability to produce compact representations of signals/images at different resolution levels. Wavelet decomposition has been used in the area of signal processing for speech compression [10]. In [10], Joseph et al. analyze the performance of different wavelets for compression purposes using a threshold to drop the wavelet coefficients with small amplitudes which are considered to be insignificant. In the area of image processing, wavelets have been used for compression [11-13], watermarking [14, 15], denoising [16, 17], detection [19, 23, 24], object recognition [18], and image classification [14-22], among others. In [18] and [19], the wavelet coefficients obtained from the discrete wavelet transform are fed into a classifier as features for object recognition and airplane detection and tracking, respectively. Wavelets have also been used to classify texture images using either energy features computed directly from the DWT coefficients [20] or a combination of wavelet statistical features and wavelet co-occurrence features such as energy, entropy or homogeneity [21]. In [22], the dependence between wavelets features from different subbands is studied to select the best set of wavelets features for classification purposes. Other features that can be used for image classification are edges which

can be detected by identifying local maximum from the wavelet transform coefficients using a threshold to remove the noise that interfere with such identification [23]. Wavelets have also been used in medical imaging applications such as detection of microcalcifications in mammograms by applying wavelet filters to remove the background noise and enhance the microcalcifications present in the mammograms [24].

2.2 OVERCOMPLETE DICTIONARY

A dictionary $\Phi \in \Re^{M \times N}$ is a collection of elementary signals called atoms, given by:

$$\Phi = \{\phi_\gamma\}_{\gamma \in \Gamma}, \quad \Gamma = \{1,2,...,M\} \tag{2.11}$$

where the atoms ϕ_γ are discrete time signals of length N. A dictionary can be classified as undercomplete, complete or overcomplete depending on whether it spans the signal space or not. If the atoms of the dictionary entirely span the signal space forming a basis, the dictionary is called a complete dictionary. When the number of atoms is larger than the dimension of the signal space $(M >> N)$ and there is a subset in the dictionary that forms a basis, it is called an overcomplete dictionary. Otherwise, the dictionary is called an undercomplete dictionary. In this case, the number of atoms composing the dictionary is less than dimension of the signal space $(M < N)$. Overcomplete dictionaries are constructed using combinations of bases or adding basis functions to a complete dictionary. The most commonly used functions to construct dictionaries are Gabor [57], wavelets [10, 11], contourlets [25], curvelets [16], ridgelets or combinations of these. Overcomplete dictionaries have become an important tool in the signal processing area due to their capacity to generate sparse representations of signals [26, 27].

A signal $\boldsymbol{y} \in \Re^N$ can be represented as a linear combination of the elements from a dictionary as follows:

$$\boldsymbol{y} = \sum_{n=0}^{N-1} \alpha_n \phi_{\gamma n} \tag{2.12}$$

where α_n are the expansion coefficients of the signal and $\gamma \in \Gamma$ is the index of the atom ϕ. However, for the case of overcomplete dictionaries, such a representation is not unique. This gives us the possibility to find the combination that works best for the desired problem. For the sparse representation problem, the goal is to find the most compact representation that reconstructs the signal with the minimum reconstruction error.

2.3 SPARSE REPRESENTATION

Given an signal $\boldsymbol{y} \in \Re^m$, an overcomplete dictionary $\Phi \in \Re^{m \times k}$ that contains *k* atoms and a vector $\boldsymbol{x} \in \Re^k$ that contains the representation coefficients of the signal ***y***, the sparse representation problem can be posed as follows:

$$\min_{\boldsymbol{x}} \|\boldsymbol{x}\|_0 \quad s.t. \quad \Phi \boldsymbol{x} = \boldsymbol{y} \tag{2.13}$$

where $\|\boldsymbol{x}\|_0$ is the $l_0 - norm$, that counts the nonzero elements in the vector $\boldsymbol{x}$. In order for the signal reconstruction to be robust to noise, equation (2.13) can be relaxed to:

$$\min_{\boldsymbol{x}} \|\boldsymbol{x}\|_0 \quad s.t. \quad \|\boldsymbol{y} - \Phi \boldsymbol{x}\|_2 \leq \varepsilon. \tag{2.14}$$

where ε is the permitted error in the reconstruction. The solution for the $l_0 - norm$ has been shown to be NP-hard. Several algorithms have been studied to obtain an approximate solution to this problem [27-32]. The two most common methods are greedy algorithms and convex optimization methods.

The convex optimization methods solve the combinatorial problem by replacing the $l_0 - norm$ with a convex function, usually with the $l_1 - norm$. In [27], Chen et al. proposed an algorithm called Basis Pursuit (BP) to solve the sparse representation problem in overcomplete dictionaries using a convex optimization method which finds the decomposition that minimizes the $l_1 - norm$:

$$\min_{\boldsymbol{x}} \|\boldsymbol{x}\|_1 \quad s.t. \quad \Phi\boldsymbol{x} = \boldsymbol{y} \tag{2.15}$$

One drawback with convex optimization methods is that they tend to be computationally extensive when the system to be solved is very large [33]. Least Absolute Shrinkage and Selection Operator (LASSO) is a method proposed by Tibshirani [34] to solve the l_1-minimization problem more efficiently. LASSO finds an estimate of $\boldsymbol{x}$ by minimizing the least square error subject to a $l_1 - norm$ constraint in the solution vector, formulated as:

$$\min_{\boldsymbol{x}} \frac{1}{2}\|\boldsymbol{y} - \Phi\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_1 \tag{2.16}$$

where $\lambda > 0$ is the parameter that controls the tradeoff between the least square error and the sparsity of $\boldsymbol{x}$. This optimization problem converges to the solution of the l_1-minimization problem when the value of λ approaches to zero.

On the other hand, greedy algorithms try to find the "best" solution to the problem iteratively. Greedy algorithms may not find the optimal solution but find a local solution that approximates to the global solution. Some algorithms proposed to solve the sparse representation problem are Matching Pursuit (MP), Orthogonal Matching Pursuit (OMP), and Compressive Sampling Pursuit (CoSaMP).

2.3.1 Matching Pursuit

Mallat and Zhang [30] introduced an algorithm, called Matching Pursuit, to represent a signal as a linear combination of elements from an overcomplete dictionary. These elements are chosen iteratively, one by one, such that they represent the structure of the signal.

Given a dictionary $\Phi = \{\phi_\gamma\}_{\gamma \in \Gamma}$ with unit norm, let $\boldsymbol{R}_n$ be the residual of an n term approximation of a given signal $\boldsymbol{y}$. MP decomposes the residue $\boldsymbol{R}_n$ by projecting it onto the elements of Φ to find the element that is highly correlated with $\boldsymbol{R}_n$. For example, the first iteration of the algorithm will represent the signal as:

$$\boldsymbol{y} = \langle \boldsymbol{y}, \phi_{\gamma 0} \rangle \phi_{\gamma 0} + \boldsymbol{R}_1 \tag{2.17}$$

where $\boldsymbol{R}_1$ is the residue after approximating $\boldsymbol{y}$ in the direction of $\phi_{\gamma 0}$. Since the norm of $\boldsymbol{y}$ can be calculated as:

$$\|\boldsymbol{y}\|_2^2 = |\langle \boldsymbol{y}, \phi_{\gamma 0} \rangle|^2 + \|\boldsymbol{R}_1\|_2^2, \tag{2.18}$$

the atom $\phi_{\gamma n} \in \Phi$ to be chosen at the n^{th} iteration is the one that solves the following optimization problem:

$$\phi_{\gamma n} = \arg\max_{\gamma \in \Gamma} |\langle \boldsymbol{R}_{n-1}, \phi_\gamma \rangle|. \tag{2.19}$$

Then, using the atom that solved the optimization method, the approximation vector α_n and the residue of the signal are updated as follows:

$$\alpha_n = \alpha_{n-1} + \langle \boldsymbol{R}_{n-1}, \phi_{\gamma n} \rangle$$

$$\boldsymbol{R}_n = \boldsymbol{R}_{n-1} - \langle \boldsymbol{R}_{n-1}, \phi_{\gamma n} \rangle \phi_{\gamma n} \tag{2.20}$$

respectively. This process is repeated until the stop criterion is met. A description of the MP algorithm can be found in Figure 5.

Input: Signal $\boldsymbol{y} \in \Re^m$, dictionary $\Phi \in \Re^{m \times k}$

Output: Sparse approximation vector α

$\boldsymbol{R}_0 \leftarrow \boldsymbol{y}$ (Residue Initialization)

$\alpha_0 \leftarrow 0$ (Initial approximation)

$n \leftarrow 1$

repeat

$\phi_{\gamma n} \leftarrow \arg\max_{\gamma \in \Gamma} \left| \left\langle \boldsymbol{R}_{n-1}, \phi_\gamma \right\rangle \right|$ (Greedy selection)

$\alpha_n \leftarrow \alpha_{n-1} + \left\langle \boldsymbol{R}_{n-1}, \phi_{\gamma n} \right\rangle$ (Approximation update)

$\boldsymbol{R}_n \leftarrow \boldsymbol{R}_{n-1} - \left\langle \boldsymbol{R}_{n-1}, \phi_{\gamma n} \right\rangle \phi_{\gamma n}$ (Residue update)

$n \leftarrow n+1$

until stop criterion

Figure 5. Matching Pursuit Algorithm

Depending on the problem to be solved, the algorithm can be stopped when one of the following stopping criteria is met: (i) after *l* fixed number of iterations, (ii) when the residue $\boldsymbol{R}_n = 0$ or (iii) when the residue $\boldsymbol{R}_n < \varepsilon$ for some ε. One shortcoming of the MP algorithm is that it requires a large number of iterations to converge, therefore it is computationally complex. Pati and Krishnaprasad [31] introduced a new algorithm called Orthogonal Matching Pursuit that overcomes this issue.

2.3.2 Orthogonal Matching Pursuit

Orthogonal Matching Pursuit, a modified version of MP, is a recursive algorithm that computes representations of functions with respect to nonorthogal and overcomplete dictionaries [31]. MP was modified by adding a least-squares minimization to improve the convergence of the algorithm.

The first step of the OMP algorithm is the same as the MP, which finds the index of the atom that is the most correlated with the residue. Then, this index is stored in the set Ω that will contain the indices of all the atoms selected through all the iterations. The third step, which is the part that is different from MP, finds the coefficient vector α_n that solves the following least-squares minimization:

$$\alpha_n = \arg\min_{\boldsymbol{x}} \left\| \boldsymbol{y} - \Phi_{|\Omega_n} \boldsymbol{x} \right\|_2 \tag{2.21}$$

where the symbol $\Phi_{|\Omega_n}$ represents the dictionary Φ with columns restricted to the indices in Ω_n. This step ensures the orthogonality between the residue and the atoms selected in the previous iterations. Therefore, the correlation of the residue with the atoms selected in the following iterations will be equal to zero avoiding the selection of the previously selected atoms, leading to a faster convergence of the algorithm. The last step updates the current residue by subtracting the projection of the signal, obtained using the restricted dictionary $\Phi_{|\Omega_n}$, from the original signal. These steps are repeated until the desired stopping criterion is met. The same stopping criteria mentioned in the previous section can be applied to the OMP algorithm. A mathematical description of the algorithm can be found in Figure 6.

Input: Signal $\boldsymbol{y} \in \Re^m$, dictionary $\Phi \in \Re^{m \times k}$

Output: Sparse coefficient vector α

$\boldsymbol{R}_n \leftarrow \boldsymbol{y}$ (Residue Initialization)

$\alpha_0 \leftarrow 0$ (Initial approximation)

$\Omega \leftarrow \emptyset$

$n \leftarrow 1$

repeat

$\gamma_n \leftarrow \arg\max_{\gamma \in \Gamma} \left| \left\langle \boldsymbol{R}_{n-1}, \phi_\gamma \right\rangle \right|$ (Atom Index)

$\Omega_n \leftarrow \Omega_{n-1} \bigcup \gamma_n$ (Atoms Index Merge)

$\alpha_n \leftarrow \arg\min_{\boldsymbol{x}} \left\| \boldsymbol{y} - \Phi_{|\Omega_n} \boldsymbol{x} \right\|_2$ (Least square minimization)

$\boldsymbol{R}_n \leftarrow \boldsymbol{y} - \Phi_{|\Omega_n} \alpha_n$ (Residue update)

$n \leftarrow n+1$

until stop criterion

Figure 6. Orthogonal Matching Pursuit Algorithm

2.3.3 Compressive Sampling Matching Pursuit

Compressive Sampling Matching Pursuit [32] is one of the modern methods used to obtain sparse representations of signals based on OMP and the compressed sensing theory. One of the differences between this algorithm and the algorithms previously described is its ability to choose multiple atoms per iteration allowing a faster convergence. CoSaMP searches iteratively for the largest elements of the target signal using a proxy signal. Needell and Tropp state that, given an *s*-sparse signal $\boldsymbol{x} \in \Re^k$ (where a signal is *s*-sparse if it has only $s < k$ nonzero elements) and a dictionary $\Phi \in \Re^{m \times k}$ whose transpose is Φ^t, a vector $\boldsymbol{p} = \Phi^t \Phi \boldsymbol{x}$ can serve as a proxy signal for the signal $\boldsymbol{x}$ due to the fact that the s largest elements in $\boldsymbol{p}$ will correspond to

the s largest elements in $\boldsymbol{x}$. Based on this, it is only necessary to apply Φ^t to the sample measurements $\boldsymbol{y} = \Phi\boldsymbol{x}$ to obtain the proxy signal.

Given a signal $\boldsymbol{y} \in \Re^m$, a dictionary $\Phi \in \Re^{m \times k}$ that meets the restricted isometric property [see 32] and a residue $\boldsymbol{r}$ initialized as $\boldsymbol{y}$, CoSaMP uses an observation vector $\boldsymbol{p} = \Phi^t \boldsymbol{r}$, to select the indices of the $2s$ atoms that are the most correlated with the residue, where *s* is the desired sparsity level. Then, this set of indices is merged with the indices of the atoms used to obtain the previous *s*-sparse coefficient (or representation) vector α_{j-1}. This new set of indices is stored in T and used to obtain the solution of the following least squares problem:

$$\boldsymbol{b}_{|T} = \Phi_{|T}^{+} \boldsymbol{y} \tag{2.22}$$

where $\boldsymbol{b}_{|T}$ is the solution of the least square problem with nonzero values corresponding to the indices in T, $\Phi_{|T}$ is the dictionary restricted to the set of indices T and $\Phi_{|T}^{+} = (\Phi_{|T}^{t} \Phi_{|T})^{-1} \Phi_{|T}^{t}$ is its pseudoinverse. The new coefficient vector α_j is obtained by retaining only the *s* largest elements in $\boldsymbol{b}_{|T}$. Finally, the residue is updated to be used in the proxy signal at the next iteration. These steps are repeated until the desired stopping criterion is met. In [32], three different stopping criteria that depend on: (i) a fixed number of iterations, (ii) the norm of the residue $\|\boldsymbol{r}\|_2$, and (iii) the maximum magnitude of the entries of the proxy $\|\boldsymbol{p}\|_\infty$ were analyzed. A list of the assumptions made and variations of this algorithm can be found in [32]. Figure 7 presents the mathematical description of the algorithm.

Input: Signal $\boldsymbol{y} \in \Re^m$, dictionary $\Phi \in \Re^{m \times k}$, sparsity level s

Output: s-sparse coefficient vector α

$\boldsymbol{r} \leftarrow \boldsymbol{y}$ (Residue initialization)

$\alpha_0 \leftarrow 0$ (Initial approximation)

$j \leftarrow 1$

repeat

$\boldsymbol{p} \leftarrow \Phi^t \boldsymbol{r}$ (Generate the proxy)

$\Omega_j \leftarrow \text{supp}(\boldsymbol{p}_{2s})$ (Index of 2s largest coeff.)

$T \leftarrow \Omega_j \bigcup \text{supp}(\alpha_{j-1})$ (Update the set of indices)

$\boldsymbol{b}_{|T} \leftarrow \Phi_{|T}^{+} \boldsymbol{y}$ (Least squares problem)

$\boldsymbol{b}_{|T^c} \leftarrow 0$

$\alpha_j \leftarrow \boldsymbol{b}_s$ (Pruning step)

$\boldsymbol{r} \leftarrow \boldsymbol{y} - \Phi \alpha_j$ (Residue update)

$j \leftarrow j + 1$

until stop criterion

Figure 7. Compressive Sampling Matching Pursuit Algorithm

2.4 COMPRESSED SENSING

Nyquist sampling theorem states that a continuous time signal can be reconstructed from its samples if the sampling rate is greater than twice the signal bandwidth. In many applications such as bioinformatics, astronomy, machine learning, hyperspectral and medical imaging, the number of samples required by the Nyquist sampling theorem can be too high which leads to the problem of having to handle high dimensionality data. In recent years, it has been shown that a sparse signal can be reconstructed from a small number of random measurements less than the minimum number of samples required by the Nyquist sampling theorem, using compressed sensing framework [35].

Given a signal $\boldsymbol{y} \in \Re^m$ that is *s*-sparse in a dictionary $\Phi \in \Re^{m \times k}$ with representation coefficients contained in the *s*-sparse signal $\boldsymbol{x} \in \Re^k$ and a sampling matrix $\Psi \in \Re^{N \times m}$, the signal $\boldsymbol{y}$ can be recovered from its measurements:

$$\boldsymbol{z} = \Psi \boldsymbol{y} = \Psi \Phi \boldsymbol{x}, \tag{2.23}$$

by solving the following optimization problem:

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{y}} \|\boldsymbol{x}\|_0 \quad \text{subject to} \quad \boldsymbol{z} = \Psi \boldsymbol{y} = \Psi \Phi \boldsymbol{x} \tag{2.24}$$

where $\Psi\Phi$ satisfies the restricted isometry property (RIP) [36, 40]. It is said that a matrix Ψ satisfies the RIP of order s if there exist a constant $\delta_s \in (0,1)$ such that

$$(1-\delta_s)\|\boldsymbol{x}\|_2^2 \le \|\Psi \boldsymbol{x}\|_2^2 \le (1+\delta_s)\|\boldsymbol{x}\|_2^2 \tag{2.25}$$

holds for all s-sparse vectors (with at most s nonzero entries). This property ensures near optimal performance of reconstruction algorithms [38]. In [39], it has been proven that random matrices whose entries are independent and identically distributed (i.i.d) random variables, such as Gaussian, Bernoulli or related distributions, will satisfy the RIP. Therefore, if the sampling matrix Ψ is chosen as one of these random matrices, there is no need to know the representation matrix Φ and it can be chosen arbitrarily [39, 40]. As mentioned, the solution to the $l_0 - norm$ in equation (2.24) it is known to be a NP-hard problem that can be solved either by greedy algorithms or convex optimization methods.

From equation (2.23), it can be seen that the sampling matrix Ψ reduces the dimension of the signal $\boldsymbol{y}$ by mapping the signal from the *m*-dimensional space to the *N*-dimensional space where N is much smaller than m. This reduction is obtained through the inner product of the random vectors of the sampling matrix with the sparse signal producing a low dimensional signal $\boldsymbol{z}$ of random measurement as shown in Figure 8.

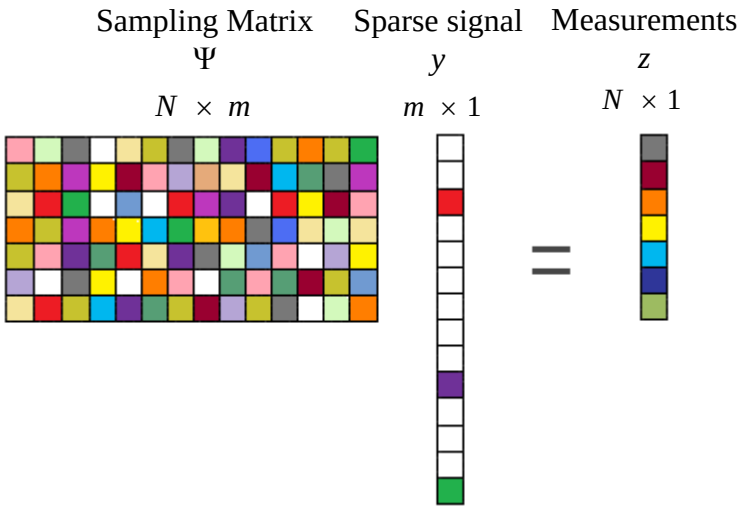

Figure 8. Compressed sensing model (For interpretation of the references to color in this and all other figures, the reader is referred to the electronic version of this thesis.)

Compressed sensing has been used in the area of medical imaging to reduce the acquisition time of MRI images by reducing the number of measurements to be acquired for reconstruction purposes [41]. It has also been used to reduce the complexity of the framework of radars by eliminating the necessity of matched filtering in the radar receiver and reducing the sampling rate of the receiver [42]. The compressed sensing theory has lead to the design of the "single pixel" compressed sensing camera that computes random linear measurements (inner products between the scene and a set of test functions) of the scene under view instead of pixel samples [43]. This camera compresses the scene under view at the same time that it is acquired, thus, it has the ability to handle high-dimensional data set.

CHAPTER 3

DISCRIMINATIVE SPARSE REPRESENTATIONS FOR IMAGE CLASSIFICATION USING PURSUIT ALGORITHMS

Sparse representations have been actively used in areas such as noise removal [44], inpainting [45], compression [46], reconstruction [30-32], among other areas of signal processing due to the fact that many natural signals can be sparse with respect to the proper dictionary. Therefore, a signal can be represented with a small number of dictionary elements such that its reconstruction error is minimal. Recently, there has been an effort to expand sparse representation to the area of image classification (e.g., [47-49]). In [47], Wright et al. uses the training images as the atoms of the dictionary to obtain the sparse representation of the test images for the purpose of face recognition. The classification is performed by assigning the test image to the class that minimizes the residual. This algorithm performs well when the objects to be classified have minimal pose variation. However, for objects with high intra-class variation, the representation may no longer be sparse. Also, it requires a large number of training images per class to be able to represent the test samples as a linear combination of only the training images from the same class. In [48], Huang and Aviyente proposed a framework that combines reconstruction error and discrimination power to obtain sparse and discriminative representation of signals using MP. A similar method can be found in [49], where a metric that includes both reconstruction and discrimination terms is proposed to learn adaptive dictionaries which leads to sparse discriminative and reconstructive image representation. This method learns one dictionary per class which can be computationally complex.

The most widely used method to obtain discriminative representations of images from different classes is Linear Discriminative Analysis (LDA). However, as mentioned before, LDA is known to be highly sensitive to the noise in data. On the other hand, the sparse representation method is robust to noise but its main objective is to look for representations that are suitable for reconstruction. To be able to classify images with high classification accuracy, it is more important to obtain discriminative representations rather than reconstructive representations. To reach this goal, a robust algorithm is proposed based on greedy pursuit algorithms that instead of choosing atoms for reconstruction purposes, chooses the atoms that are suitable for image classification.

3.1 DISCRIMINATIVE SPARSE REPRESENTATIONS

In this chapter, an objective function that combines discrimination power and sparsity to obtain discriminative representations of images is introduced. To demonstrate the effectiveness of using discriminative representations instead of reconstructive representations, the original CoSaMP was modified so that classification results can be obtained using reconstructive features as well as discriminative features. Then, an algorithm derived from CoSaMP, in conjunction with the objective function, is proposed to select the smallest possible number of atoms that produce the best discriminative representation of a set of images. This algorithm was inspired by CoSaMP and the simultaneous sparse approximation algorithms described in [50, 51]. Finally, some classification experiments that demonstrate the effectiveness of the objective function and the robustness of the proposed algorithm is presented.

3.1.1 Objective Function

Given an overcomplete dictionary $\Phi \in \Re^{m \times k}$ with k atoms of dimensionality m, an input matrix $\boldsymbol{Y} = [\boldsymbol{Y}_1, \boldsymbol{Y}_2, ..., \boldsymbol{Y}_c]$ where $\boldsymbol{Y}_i = [\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_{n_i}]_{1 \le i \le c}$ corresponds to the set of n_i vectorized training images from the i^{th} class given that there are $C = \{1,2,...,c\}$ different classes and $n = \sum_{i=1}^{c} n_i$ is the total number of images, the feature matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n]$ can be obtained by projecting the training images against the atoms in the dictionary where $\boldsymbol{x}_j$ is the k-dimensional feature vector that corresponds to the j^{th} training image. To measure how well each atom in the dictionary produce a discriminative representation of the n training images, the following discrimination measure is used:

$$F(\boldsymbol{X}) = \frac{trace(\boldsymbol{S}_b)}{trace(\boldsymbol{S}_w)}, \tag{3.1}$$

where $\boldsymbol{S}_b$ and $\boldsymbol{S}_w$ are the between-class and the within-class scatter matrices, respectively defined as:

$$\boldsymbol{S}_b = \sum_{i=1}^{c} n_i (\boldsymbol{\mu}_i - \mu)(\boldsymbol{\mu}_i - \mu)^t \tag{3.2}$$

$$\boldsymbol{S}_w = \sum_{i=1}^{c} \sum_{x_j \in C_i} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^t \tag{3.3}$$

where,

$$\boldsymbol{\mu}_i = \frac{1}{n_i} \sum_{\boldsymbol{x}_i \in C_i} \boldsymbol{x}_i \tag{3.4}$$

and μ is the total mean vector defined as:

$$\mu = \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{x}_j \tag{3.5}$$

This measurement *F(X)* will produce a vector of size $1 \times k$ that contains the discriminative power of each atom. The highest values in this vector correspond to the atoms that produce a discriminative representation of the training images by maximizing the between-class scatter and minimizing the within-class scatter. Given that $\tilde{\boldsymbol{X}} \in \Re^{s \times n}$ is a subset of $\boldsymbol{X} \in \Re^{k \times n}$ ($s < k$), the combination of discrimination and sparsity measures is proposed to identify the atoms that discriminate and create a sparse representation of the training signals by maximizing the following objective function:

$$\tilde{\boldsymbol{X}} = \arg\max_{\boldsymbol{X}} F(\boldsymbol{X}) - \left\| \sum_{i=1}^{n} \boldsymbol{x}_i \right\|_0 \tag{3.6}$$

where the rows of the subset matrix $\tilde{\boldsymbol{X}}$ are the most discriminative representations. The indices of the atoms that produce these discriminative representations will be stored in the set Ω to obtain the low dimensional features of the test images.

3.1.2 Discriminative CoSaMP vs. Reconstructive CoSaMP

In the literature, it has been shown that discriminative representations perform better than reconstructive representations for classification purposes [68, 69]. To illustrate that this is also true for the case of sparse representations, CoSaMP was used to evaluate the effect of selecting atoms that produce discriminative representations of the images versus atoms that produce reconstructive representations. The original CoSaMP algorithm was modified to be able to obtain the atoms that simultaneously produce a reconstructive representation of a set of images from

different classes. Then, the objective function (3.6) was introduced to the algorithm to obtain the atoms that produce a discriminative representation of the same set of images.

Given a signal $\boldsymbol{Y} \in \Re^{m \times n}$ and a dictionary $\Phi \in \Re^{m \times k}$ with *k* atoms, the reconstructive CoSaMP (RecCoSaMP) algorithm starts by initializing the residue $\boldsymbol{R}$ as the input matrix $\boldsymbol{Y}$ and the representation matrix $\boldsymbol{A} \in \Re^{k \times n}$ as a matrix of zeros. Then, the algorithm obtains the set of indices of the atoms that produce a representative representation of the images in $\boldsymbol{Y}$ by:

(1) Generating a proxy matrix $\boldsymbol{P}(r,c) = \Phi^t \boldsymbol{R}$ where $c = \{1,2...,n\}$ and $r = \{1,2...,k\}$.

(2) Obtaining the largest coefficient from each row of the proxy matrix and storing it in the vector $\boldsymbol{p}(r)$.

(3) Identifying the indices of the 2s largest coefficients in the vector $\boldsymbol{p}(r)$.

(4) Merging the identified indices in step (3) with the set used to obtain the previous representation matrix $\boldsymbol{A}$. At the first iteration, the set used to obtain the representation matrix is empty.

(5) Solving the least squares problem by restricting the columns in the dictionary using the indices on the merged set.

(6) Obtaining the largest coefficient from each row of the approximation matrix obtained in (5) and storing it in the vector $\boldsymbol{b}(r)$.

(7) Identifying the indices of the *s* largest coefficients in the vector $\boldsymbol{b}(r)$.

(8) Retaining only the dimensions of the approximation matrix obtained in step (5) using the indices identified in step (7) and setting all other dimensions to zero to obtain the *s*-sparse representation matrix $\boldsymbol{A}$.

(9) Updating the residue by subtracting the part of the input matrix that has been already approximated from the input matrix.

These steps are repeated until the stop criterion is met. For the Discriminative CoSaMP (DiscCoSaMP) algorithm, the first five steps (1)-(5) are similar to the RecCoSaMP algorithm. Then, the objective function (3.6) is introduced to identify the s indices of the atoms that produce a discriminative representation of the input images. This set of indices is used to retain only the most significant dimensions of the representation matrix $\boldsymbol{B}$ and setting all other dimensions to zero. Finally, similar to the RecCoSaMP algorithm, the residue is updated. A mathematical description of these algorithms can be found in Figure 9.

In the first four steps of the DiscCoSaMP, a set of $2s$ atoms are selected based on the magnitude of the coefficient of the proxy signal to obtain the result of the least squares problem. These steps, however, may be redundant since this set of atoms is pruned to obtain only the s indices of the most discriminative atoms. The indices of atoms that produce a discriminative representation of the images can be selected directly from the solution of the least squares problem while still obtaining sparse representations with high discrimination power for image classification.

Input: Signal $\boldsymbol{Y} \in \Re^{m \times n}$, dictionary $\Phi \in \Re^{m \times k}$, sparsity level s
Output: Set of indices Γ

Reconstructive CoSaMP	Discriminative CoSaMP
$\boldsymbol{R} \leftarrow \boldsymbol{Y}$	$\boldsymbol{R} \leftarrow \boldsymbol{Y}$ (Residue initialization)
$\boldsymbol{A}_0 \leftarrow 0$	$\boldsymbol{A}_0 \leftarrow 0$ (Initial approximation)
$\Gamma_0 \leftarrow \emptyset$	$\Gamma_0 \leftarrow \emptyset$
$j \leftarrow 1$	$j \leftarrow 1$
repeat	**repeat**
$\boldsymbol{P}(r,c) \leftarrow \Phi^t \boldsymbol{R}$	$\boldsymbol{P}(r,c) \leftarrow \Phi^t \boldsymbol{R}$ (Generate the proxy)
$\boldsymbol{p}(r) \leftarrow \max \boldsymbol{P}(r,c)$	$\boldsymbol{p}(r) \leftarrow \max \boldsymbol{P}(r,c)$ (Largest coeff. in each row)
$\Omega \leftarrow \arg\max_r (\boldsymbol{p}(r)_{2s})$	$\Omega \leftarrow \arg\max_r (\boldsymbol{p}(r)_{2s})$ (Index of 2s largest coeff.)
$T \leftarrow \Omega \bigcup \Gamma_{j-1}$	$T \leftarrow \Omega \bigcup \Gamma_{j-1}$ (Update the set of indices)
$\boldsymbol{B} \leftarrow \Phi_{\|T}^{+} \boldsymbol{Y}$	$\boldsymbol{B} \leftarrow \Phi_{\|T}^{+} \boldsymbol{Y}$ (Least squares problem)
$\boldsymbol{b}(r) \leftarrow \max \boldsymbol{B}(r,c)$	$\Gamma_j \leftarrow \arg\max_{\boldsymbol{B}} F(\boldsymbol{B})$
$\Gamma_j \leftarrow \arg\max_r \boldsymbol{b}(r)$	$\boldsymbol{A}_j \leftarrow \boldsymbol{B}_{\|\Gamma}$ (Pruning step)
$\boldsymbol{A}_j \leftarrow \boldsymbol{B}_{\|\Gamma}$	$\boldsymbol{R} \leftarrow \boldsymbol{Y} - \left\|\Phi \boldsymbol{A}_j\right\|$ (Residue update)
$\boldsymbol{R} \leftarrow \boldsymbol{Y} - \left\|\Phi \boldsymbol{A}_j\right\|$	$j \leftarrow j+1$
$j \leftarrow j+1$	**until** stop criterion
until stop criterion	

Figure 9. Reconstructive CoSaMP and Discriminative CoSaMP algorithms

3.1.3 Discriminative Sparse Representations Algorithm

In this section, the proposed method derived from CoSaMP to obtain the indices of the atoms that produce a discriminative representation of a set of input images from different classes is presented [52]. As input, the proposed algorithm needs a matrix containing the m-dimensional vectorized training images $\boldsymbol{Y} \in \Re^{m \times n}$, the overcomple dictionary $\Phi^{m \times k}$ with k atoms, the stop criterion and the desired sparsity level s (the number of atoms to obtain the discriminative

representation of the images). Similar to the greedy methods explained in Chapter 2, the residue $\boldsymbol{R}$ is initialized as the input matrix $\boldsymbol{Y}$. Then, the proposed algorithm selects the indices of the atoms that produce a discriminative sparse representation of the input matrix by:

(1) Solving the least square problem, $\boldsymbol{X} = \Phi^{+}\boldsymbol{R}$, to obtain the approximation matrix.

(2) Getting the set Ω of s indices of the atoms that solve the optimization problem (3.6) using the approximation matrix from step (1).

(3) Pruning the approximation matrix by retaining only the rows produced by the atoms in the set of indices obtained in (2) and setting all other rows to zero.

(4) Updating the residue through the subtraction between the input matrix and the part of the signal that has been approximated.

These steps are repeated until the stopping criterion selected is met. A mathematical description of the proposed algorithm can be found in Figure 10.

Input: Matrix of images $\boldsymbol{Y}^{m\times n}$, dictionary $\Phi^{m\times k}$, sparsity level s
Output: Dictionary indices Ω

$\boldsymbol{R}_0 \leftarrow \boldsymbol{Y}$ (Residue initialization)
$l \leftarrow 1$
repeat
  $\boldsymbol{X} \leftarrow \Phi^{+}\boldsymbol{R}_{l-1}$ (Least squares problem)
  $\Omega_l \leftarrow \arg\max_{\boldsymbol{X}} F(\boldsymbol{X})$ (Index of s largest values)
  $\boldsymbol{A}_s \leftarrow \boldsymbol{X}_{|\Omega_l}$ (Sparse Approximation)
  $\boldsymbol{R}_l \leftarrow \boldsymbol{Y} - |\Phi \boldsymbol{A}_s|$ (Residue update)
  $l \leftarrow l + 1$
until stop criterion

Figure 10. Discriminative Sparse Representations Algorithm

In the literature, several methods are described to solve the least squares problem. In [32], the use of Richardson's iteration or conjugate gradient method is recommended to solve this problem. In this thesis, the biconjugate gradient stabilized method (BiCGStab) was used, which is a variant of the conjugate gradient method but have a smoother and faster convergence. As stopping criterion, given that Ω_l is the set of indices identified on the l^{th} iteration and Ω_{l-1} the set of indices identified on the $(l-1)^{th}$ iteration, the cardinality of the intersection between Ω_l and Ω_{l-1} was used. If the cardinality of the intersection is less than the pre-determined sparsity value *s*, the algorithm continues to the next iteration until the following criteria is met:

$$\left|\Omega_{l-1} \cap \Omega_l\right| = s \tag{3.7}$$

3.2 EXPERIMENTS AND RESULTS

In this section, experimental results on two different databases are presented to evaluate the performance of the algorithm. The algorithm is implemented for the cases of noiseless images and images with different levels of noise and occlusion to demonstrate the robustness of the algorithm. The experiment in Section 3.2.3 uses a subset of the COIL database to evaluate the performance of the proposed algorithm with images with low intra-class variability. Also, a comparison between the results of the proposed algorithm with those of a modified version of OMP will be presented to evaluate if the number of atoms selected at each iteration affects on the classification results. The classification results were obtained using Support Vector Machine (SVM) as the classifier, where the accuracy of SVM depends on the model parameters chosen to classify the images. In the second experiment, to avoid the dependency of the classification results on the classifier parameters the performance of the proposed algorithm was evaluated

using a similarity metric based on correlation. The database used in this experiment contains images with high intra-class variability to prove that the proposed algorithm can handle class variability and the results compared with those of LDA.

3.2.1 Databases and Experimental Setup

The first database used to evaluate the algorithm is the Columbia Object Image Library (COIL-20) dataset. This database consists of 1440 grayscale images of 20 different objects as shown in Figure 11. Each object was placed in a motorized turntable against a black background and it was rotated through 360 degrees to obtain images every 5 degrees, for a total of 72 images per object [53]. Each image in the database was resized from 128×128 pixels to 16×16 pixels and rearranged into a vector of length 256. These vectors were divided into training and testing sets and the vectors of each set were concatenated in a matrix where the training matrix was used to obtain the atoms that simultaneously represent the images and the test matrix to obtain the classification accuracies. From this database only a subset of 423 images containing the first six objects was used, which is going to be called Coil-1.

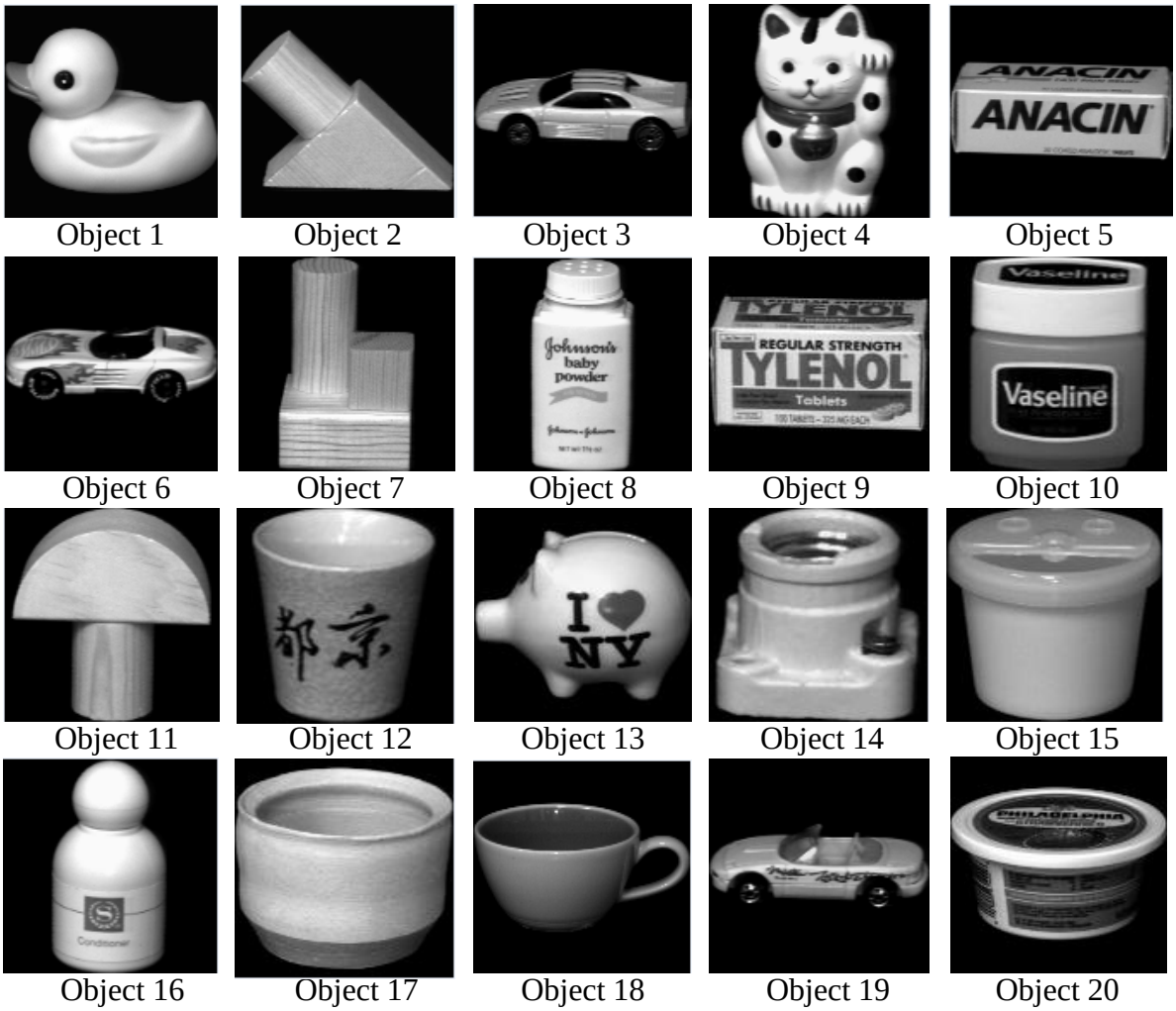

Figure 11. Sample images from COIL-20 database (The images in this figure are for visual reference only, the text in the images are not meant to be readable)

The second database is a more challenging database whose objects are against real world backgrounds, under different lightning conditions, with high intra-class variability and can even contain some occlusions. The database is the TU Darmstadt Database from the PASCAL Object Recognition Database Collection, formerly the ETHZ database [54]. This database consists of images from three different classes: side views of motorbikes, cows and cars. Samples of the images from this database can be seen in Figure 12. From this database a subset of 50 images per class was extracted for a total of 150 images. The objects in the images were extracted with the

segmentation mask provided and resized into 128×128 images, whose pixels were rearranged into a vector of length 16,384. Similar to the previous database, the vectors were divided into training and testing sets and the vectors of each set were concatenated in a matrix where the training matrix was used to obtain the atoms that simultaneously represent the images and the test matrix to obtain the classification accuracies.

Figure 12. Sample images from the ETHZ database

The robustness of the algorithms was evaluated by applying different levels of noise and occlusion to the images. For both databases, the noisy images were generated using random Gaussian noise with signal-to-noise ratios (SNR) of 20dB, 15dB and 10dB. The occluded images from the Coil database contain black squares of size 3×3, 5×5 and 7×7. For the case of the images from the ETHZ database, the black squares are of size 15×15, 19×19 and 31×31.

In both cases, the features of the images were extracted by using an overcomplete dictionary formed by a combination of Haar atoms and Gabor functions. Haar wavelets have been used to detect/model discontinuities in images [55, 56]. On the other hand, Gabor functions are good for modeling continuous elements and directionality in images. Some advantages of the Gabor functions are their ability to save the neighborhood relation between pixels, its robustness against illumination and noise and its ability to represent images based on the way the human mind does [57]. The ability of these bases to model both continuities and discontinuities make them suitable to classify natural images that consist of continuous and discontinuous elements. The overcomplete dictionary used in the experiments contains 256 Haar atoms and 781 Gabor atoms for a total of 1037 atoms. The Haar atoms were generated using the scaling function corresponding to the length of the vectorized images and the wavelet functions for all shifts and eight scales. And the Gabor functions were generated as:

$$G(x) = g(x)\cos(w_0(x - x_0)) \tag{3.8}$$

where $w_0 = 5$ is the center of support in the frequency domain and

$$g(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\left(\frac{x - x_0}{\sigma}\right)^2}{2}\right) \tag{3.9}$$

is a Gaussian function with scale $\sigma = 0.05$ and center x_0.

3.2.2 Reconstructive vs. Discriminative Representation

The first experiment evaluates the performance of the RecCoSaMP and DiscCoSaMP to illustrate how the inclusion of a cost function that quantifies separability between the images' classes can improve the classification accuracy. This experiment was performed using Coil-1,

which contains 432 images from 6 different classes, was performed to evaluate the performance of the Reconstructive and Discriminative CoSaMP algorithms presented in Section 3.1.2. The experiment was performed using 396 images as training images and 36 as test images. For the classification stage, SVM was used as the classifier and the results are an average of a 12-fold cross-validation.

SVM is a method that uses a kernel mapping function which transforms the input data to a higher dimensional plane seeking the optimal hyperplane that best separates the feature vectors from different classes. Some of the most commonly used kernel functions are lineal, polynomial, radial basis functions and sigmoids. SVM was implemented using the library LibSVM [58] and the radial basis function as the kernel function. The accuracy of the algorithm has a large dependency on the model parameters. If these parameters are not selected correctly, the accuracy results may not be optimal. In [59], it is recommended to use the grid search to find the optimal values of the parameters. This is a method that evaluates the performance of the algorithm trying different values of each parameter across a desired range. Then, the set of parameters with the "best" accuracy is picked.

The goal of this work is to obtain the maximum classification accuracy with the smallest number of atoms. Therefore, the results are evaluated in terms of the optimal and maximum results using the following cost function:

$$\max(\alpha A - \beta S_p) \tag{3.10}$$

where A is the percentage of classification, S_p is the percentage of sparsity (number of atoms selected to obtain the accuracy/total number of atoms in the dictionary) and α, β are constants. When $\alpha = 1$ and $\beta = 1$, the optimal results which correspond to the point where there is an optimal tradeoff between the accuracy and the number of atoms used to classify the images can

be obtained. The maximum accuracy results correspond to $\alpha = 1$ and $\beta = 0$ which is obtained when the maximum accuracy has occurred for the smallest number of atoms.

From Table 1, it can be seen that there is not a big difference between the optimal accuracies of both algorithms but the optimal sparsity of the Discriminative CoSaMP (DiscCoSaMP) is much smaller than that of the Reconstructive CoSaMP (RecCoSaMP). This is because the first atoms selected using RecCoSaMP are suitable to reconstruct the images and not necessarily to classify them, making it necessary the use of more atoms to obtain high classification results. Comparing the two algorithms, there is a difference of 6.26% in sparsity which corresponds to 65 atoms more atoms needed to obtain similar results. These results prove that using the cost function (3.6), classification accuracies near the maximum results can be obtained using much less atoms than using a reconstructive cost function.

Algorithm	Opt. / Max. Results	Noiseless (%)
Reconstructive CoSaMP	Opt. Spar.	13.91
	Opt. Max.	89.35
	Max. Spar.	21.33
	Max. Acc.	93.52
Discriminative CoSaMP	Opt. Spar.	7.65
	Opt. Max.	90.28
	Max. Spar.	21.48
	Max. Acc.	93.98

Table 1. Classification results using DiscCoSaMP and RecCoSaMP from Coil-1

3.2.3 Robustness of the Discriminative Sparse Representations Algorithm

In this section, an evaluation of the proposed algorithm using images with low intra-class variability is performed. The robustness and sparsity of the algorithm are evaluated with images under different levels of Gaussian noise and occlusion. Also, a comparison between the classification results obtained using the atoms selected with OMP and the proposed algorithm will be presented. The image database used was Coil-1 which contains 432 images from 6 classes with low intra-class variability. From the 432 images in the database, 396 images were used as training images to obtain the indices of the atoms that best discriminate between classes and the other 36 were used as testing images to evaluate the performance of the algorithm using SVM as the classifier. The results are the product of an average of a 12-fold cross-validation.

To be able to compare the results of the proposed algorithm with OMP, it is necessary to modify OMP in a way that it selects the atoms that produce the most discriminative representation instead of a reconstructive one. The modification was done by replacing the identification step in the OMP algorithm with the cost function in equation (3.6). The original step selected the atoms most correlated with the residue. With the introduction of the cost function (3.6), the algorithm will select the atom that produces the most sparse and discriminative representation. The mathematical description of the modified OMP algorithm can be seen in Figure 13.

Input: Signal $\boldsymbol{Y}$, dictionary Φ

Output: Set of indices Ω

$\boldsymbol{R}^0 \leftarrow \boldsymbol{Y}$ (Residue Initialization)

$\boldsymbol{A}_0 \leftarrow 0$ (Initial approximation)

$\Omega \leftarrow \emptyset$

$n \leftarrow 1$

repeat

$\boldsymbol{C} \leftarrow \left| \left\langle \boldsymbol{R}^{n-1}, \phi_\gamma \right\rangle \right|$

$\gamma_n \leftarrow \arg\max_{\gamma \in \Gamma} F(C)$ (Atom Selection)

$\Omega_n \leftarrow \Omega_{n-1} \bigcup \gamma_n$ (Merging the indices of the selected atoms)

$\boldsymbol{A}_n \leftarrow \arg\min_{\boldsymbol{X}} \left\| \boldsymbol{Y} - \Phi_{|\Omega_n} \boldsymbol{X} \right\|_2$ (Least square minimization)

$\boldsymbol{R}^n \leftarrow \boldsymbol{Y} - \left| \Phi_{|\Omega_n} \boldsymbol{A}_n \right|$ (Residue update)

$n \leftarrow n+1$

until stop criterion

Figure 13. A modified version of OMP by adding a cost function that quantifies discrimination power and level of sparsity

Table 2 shows the optimal and maximum accuracy results obtained from the modified OMP and the proposed algorithm for the case of Gaussian noise. It can be seen that, when the SNR value is low (20dB and 15dB), the optimal accuracy results from the proposed method are slightly higher compared to the modified OMP results with a maximum difference in the optimal sparsity values of around 0.58% (which corresponds to a difference of 6 atoms). When the noise level is higher, in this case 10dB, the optimal accuracy result from the proposed method is higher than the modified OMP with a difference of 0.16% in the optimal sparsity values (which corresponds to a difference of 2 atoms). For the case of images with occlusion, from Table 3, it can be seen that the proposed method has better optimal and maximum accuracy results than the

modified OMP. For the case of images with occlusion of size 7×7, the classification accuracies of the proposed method are achieved with a smaller amount of atoms than the modified OMP.

In conclusion, even though the sparsity values of both algorithms are not very different, the proposed algorithm still obtains better accuracy results than the modified OMP. These results show the robustness of the algorithm under noisy conditions and missing data. In real problems, due to the fact that the amount of noise or missing data in the images is unknown, it is better to use the proposed algorithm since higher classification results can be obtained at the cost of using only a few more atoms than the modified OMP. Also, the ability of the proposed algorithm to select multiple atoms at each iteration makes the computational time to be much faster than the modified OMP.

Method	Opt. / Max. Results	Noiseless (%)	20dB (%)	15dB (%)	10dB (%)
Modified OMP	Opt. Sparsity	3.52	3.55	3.53	4.32
	Opt. Accuracy	90.74	86.40	84.55	77.96
	Max. Sparsity	5.95	7.53	8.79	11.35
	Max. Accuracy	92.59	88.24	86.65	81.50
Proposed Algorithm	Opt. Sparsity	3.42	3.94	4.11	4.48
	Opt. Accuracy	90.74	86.73	84.80	80.71
	Max. Sparsity	4.97	4.42	5.25	6.31
	Max. Accuracy	91.44	87.05	85.20	81.15

Table 2. Classifications results using a modified OMP algorithm and the proposed greedy algorithm from Coil-1 with Gaussian noise

Method	Opt. / Max. Results	Noiseless	3×3	5×5	7×7
Modified OMP	Opt. Sparsity	3.52	3.79	3.37	3.53
	Opt. Accuracy	90.74	81.59	73.72	62.70
	Max. Sparsity	5.95	10.57	11.67	13.81
	Max. Accuracy	92.59	84.24	77.02	66.65
Proposed Algorithm	Opt. Sparsity	3.42	3.70	3.51	2.66
	Opt. Accuracy	90.74	84.31	76.87	68.75
	Max. Sparsity	4.97	5.49	5.72	4.11
	Max. Accuracy	91.44	84.92	77.58	68.12

Table 3. Classification results using a modified OMP algorithm and the proposed greedy algorithm from Coil-1 with occlusion

3.2.4 Comparison with LDA

Since LDA is one of the most commonly used methods to obtain image representation in a lower dimensionality space, the results obtained using this method were compared with those of the proposed method. The comparison was done using the ETHZ database and 10-fold cross-validation. From the 150 images, 135 were used as training images to obtain the indices of the atoms that best discriminate between classes and the other 15 images were used as test images to evaluate the performance of the algorithm. Given that the dimensionality of the images in this database is 16,383 (128×128), to obtain a representation of the images, it is necessary to construct a dictionary with the same dimensionality of the images. To avoid the construction of such a high dimensional dictionary, which would increase the computational complexity of the algorithm, nonoverlapping patches of 16×16 pixels were extracted from the images. This

produces a total of 64 nonoverlapping patches per images from which only 10, selected randomly, were used to extract the features. To have a fair comparison, the patches of the images were projected to the same dictionary used to obtain the representation of the images in the proposed algorithm. In that way, the features vectors are extracted from the same feature space.

Due to the fact that the search of the SVM parameters can be computationally expensive and it could be that the range chosen is not the one that contains the optimal values, the evaluation for this experiment was performed using a metric based on correlation which is independent of the classifier. First, the features were extracted by projecting the extracted patches to the selected atoms. Then, the l^{th} test patch is assigned to the class of the training patch that maximizes the following cost function:

$$\hat{i} = \arg\max_{i} \frac{1}{q-1} \frac{(\boldsymbol{b}_{ji} - \bar{\boldsymbol{b}}_{ji})(\boldsymbol{d}_l - \bar{\boldsymbol{d}}_l)^t}{\sigma_{\boldsymbol{b}_{ji}} \sigma_{\boldsymbol{d}_l}} \tag{3.11}$$

where

- $\boldsymbol{b}_{ji}$ is the feature vector extracted from the j^{th} patch that belongs to the i^{th} training image
- $\bar{\boldsymbol{b}}_{ji}$ is the mean of $\boldsymbol{b}_{ji}$ and $\sigma_{\boldsymbol{b}_{ji}}$ its standard deviation
- $\boldsymbol{d}_l$ is the feature vector extracted from the l^{th} testing patch from the test image
- $\bar{\boldsymbol{d}}_l$ is the mean of $\boldsymbol{d}_l$ and $\sigma_{\boldsymbol{d}_l}$ its standard deviation
- q is the dimension of the feature vector

After classifying all the patches, the test image is assigned to the class that has the maximum number of patches assigned to that class:

$$\hat{i} = \max_{i \in C} \boldsymbol{p}(i) \tag{3.12}$$

where $\boldsymbol{p}(i)$ is the vector that contains the number of test patches assigned to the i^{th} class and $\hat{i}$ is the label assigned to the test image. Finally, the classification accuracy is defined as:

$$Accuracy = \left(1 - \frac{\sum_{l=1}^{z} \boldsymbol{e}_l}{z} \right) \times 100 \tag{3.13}$$

where z is the total number of test images and $\boldsymbol{e}_l = 1$ if the test image is misclassified and $\boldsymbol{e}_l = 0$, otherwise. To avoid overfitting of the results, an n-fold cross-validation was performed for all the experiments and for the case of noisy images, the process of adding Gaussian noise to the test images on each cross-validation was performed 15 times.

The optimal and maximum accuracy results of the LDA and the proposed algorithm can be found in Table 4 and Table 5. These results show that the proposed algorithm always surpasses LDA. For the case where the images have Gaussian noise, as expected, LDA has a low performance due to its sensitivity to noise. On the other hand, the proposed method is robust to noise. Even when the images have SNR = 10 dB, the optimal classification accuracy does not get lower than 88% of accuracy with a sparsity level of, at most, 4.95% which corresponds to around 52 atoms. This is a reduction of 80% in the dimensionality of the patches, which corresponds to a reduction of 96.82% in the dimensionality of the images (520 features per images/16,384). For the case of images with occlusion, the optimal accuracies do not get lower than 89% with, at most, 42 atoms. These results show that even with images with high intra-class variability the algorithm is robust to noise and the sparsity level needed to obtain the optimal accuracy results do not get higher than 5% of the dimensionality of the images.

Method	Opt. / Max. Results	Noiseless (%)	20dB (%)	15dB (%)	10dB (%)
LDA	Opt. Sparsity	4.61	6.38	6.01	6.37
	Opt. Accuracy	81.33	65.80	63.07	62.43
	Max. Sparsity	7.73	9.01	9.79	9.20
	Max. Accuracy	82.67	67.60	65.40	64.00
Proposed Algorithm	Opt. Sparsity	2.87	4.12	4.34	4.95
	Opt. Accuracy	90.67	89.27	89.20	88.00
	Max. Sparsity	7.40	15.14	13.87	14.94
	Max. Accuracy	91.33	91.33	92.40	91.13

Table 4. Classifications results using LDA and the proposed greedy algorithm from the ETHZ database with Gaussian noise

Method	Opt. / Max. Results	Noiseless (%)	15×15 (%)	19×19 (%)	31×31 (%)
LDA	Opt. Sparsity	4.61	5.18	4.71	4.67
	Opt. Accuracy	81.33	79.87	78.87	78.13
	Max. Sparsity	7.73	7.70	7.11	7.05
	Max. Accuracy	82.67	81.33	80.27	79.47
Proposed Algorithm	Opt. Sparsity	2.87	3.16	3.49	4.05
	Opt. Accuracy	90.67	90.13	90.13	89.33
	Max. Sparsity	7.40	11.07	12.61	14.46
	Max. Accuracy	91.33	92.40	92.93	92.53

Table 5. Classification results using LDA and the proposed greedy algorithm from the ETHZ database with occlusion

3.3 CONCLUSIONS

In this chapter, a modified cost function that achieves a tradeoff between discrimination power and the sparsity was proposed and a corresponding greedy algorithm based on CoSaMP was developed to obtain the atoms that produce a discriminative sparse representation of a set of images. The atoms selected from the overcomplete dictionary based on the training images were used to extract features and classify images from the testing set. It was shown that using a cost function to selects atoms that produce discriminative representations rather than reconstructive ones can reduce the number of atoms needed to obtain high classification accuracies.

Two different experiments were performed to evaluate the performance and robustness of the proposed algorithm. To evaluate the robustness of the algorithm, different levels of Gaussian noise and occlusion were added to the test images. The experiments showed that the proposed algorithm can work under conditions where the images contain high level of noise and occlusion and still maintain a low sparsity level. Also, it was shown that the algorithm is able to handle image datasets with low intra-class variability as well as high intra-class variability. In addition, it was shown that the proposed algorithm can work with reduced dimensionality data, such as nonoverlapping random patches, instead of the whole images avoiding the construction of high dimensional dictionaries.

CHAPTER 4

DISCRIMINATIVE FEATURE SELECTION FROM COMPRESSED SENSING MEASUREMENTS FOR IMAGE CLASSIFICATION

In recent years, there has been a move in signal processing to sense signals using fewer samples than the Nyquist rate of samples, which is known as compressed sensing. Recently, this idea has been explored in the area of signal classification and detection where the high dimensionality of the data can increase the complexity of such applications. For example, the current methods to obtain images for MRI applications can be time-consuming due to the high dimensionality of the images. To solve this problem, compressed sensing has been used to improve the acquisition time of the images without degrading the image quality [60]. In [61], the use of random measurements from local patches is proposed to classify texture images using a single nearest neighbor classifier. First, a texton dictionary is learned from the compressed sensing measurements of the patches using K-means clustering and a histogram per class is learned by labeling each of the patches to the closest texton in the dictionary. Then, a new texture image is classified by finding the histogram that is closest to the new texture image histogram. In [62], compressed sensing has been used for signal detection by applying the MP algorithm to select fewer numbers of incoherent measurements than the necessary for reconstruction purposes. Compressed sensing has also been used for signal classification [63] by projecting the training signals and the test signal to be classified onto random vectors and classifying the signal to the class of the training signal with minimum distance between the compressed measurements of the test signal and the compressed measurements of the training signal.

In this thesis, the work of Haupt et al. [63] will be used as motivation for the work presented in this chapter. Given a test signal $\boldsymbol{f}_T \in \Re^n$ and a sampling matrix $\boldsymbol{A} \in \Re^{\boldsymbol{k}\times\boldsymbol{n}}$ where $\boldsymbol{A}_j$ corresponds to the element/row j in the matrix, Haupt et al. have shown that the compressed measurements of the signal $\boldsymbol{f}_T$, besides being used for reconstruction purposes, can be used as features for classification purposes. The compressed measurements of the signal $\boldsymbol{f}_T$ are obtained through the following inner product:

$$\boldsymbol{y}(j) = \left\langle \boldsymbol{A}_j, \boldsymbol{f}_T \right\rangle + \boldsymbol{w}(j) \qquad \text{for } j = 1,2,...,k \tag{4.1}$$

where $\boldsymbol{w}(j)$ is the vector that models the measurement noise with entries are i.i.d. $N(0,\sigma^2)$ random variables and $k < n$. Given a set $\boldsymbol{F} = \{\boldsymbol{f}_1, \boldsymbol{f}_2, ..., \boldsymbol{f}_m\}$ of m training signals, the classification is performed by assigning to the test signal the class of the training signal that solves the following minimization problem

$$\hat{\boldsymbol{f}}_k = \arg\min_{\boldsymbol{f}\in\boldsymbol{F}} \left\| \boldsymbol{y} - \alpha \boldsymbol{A}\boldsymbol{f} \right\|^2. \tag{4.2}$$

Haupt et al. performed several simulations using this method to validate the following theoretical classification error bound:

$$P\left(\hat{\boldsymbol{f}}_k \neq \boldsymbol{f}_T\right) \leq (m-1)\left(1 + \frac{\alpha^2 d_{\min}}{4n\sigma^2}\right)^{-k/2} \tag{4.3}$$

where

$$d_{\min} = \min_{\boldsymbol{f}_i, \boldsymbol{f}_j \in \boldsymbol{F}, i\neq j} \left\| \boldsymbol{f}_i - \boldsymbol{f}_j \right\|^2 \tag{4.4}$$

is the minimum Euclidean distance between every pair of training signals and $\alpha \in \Re$ is used for the cases where the training signals in $\boldsymbol{F}$ do not have unit norm. This method uses all of the

measurements obtained from compressed sensing without taking into account the presence of possible redundancy or irrelevant features in the measurements which may reduce the classification accuracy. This problem can be solved by applying a feature selection method in conjunction with compressed sensing to keep only the relevant features for classification purposes.

Feature selection is an important step for the area of image classification that selects a subset of relevant features from an input set of images to construct a robust model for classification purposes. Given a set of features, the elimination of irrelevant features will help to improve the classification accuracy, rather than using the whole set of features. In [64], it is shown that classification accuracy can drop significantly in the presence of irrelevant features if the feature selection method used fails to identify them. This chapter proposes to include a cost function to select the most relevant features from the compressed sensing measurements.

4.1 DISCRIMINATIVE COMPRESSED MEASUREMENTS

To select the set of measurements/features that are the most relevant for classification purposes, a metric that measures the discriminative power of each feature vector is proposed. This measurement will be used to sort the feature vectors from the random measurements and select the most discriminative features for classification purposes. Using this method, it is expected to obtain better classification accuracy using less feature vectors than using the whole set obtained with compressed sensing.

Given a set $\boldsymbol{F} = \{\boldsymbol{f}_1, \boldsymbol{f}_2, ..., \boldsymbol{f}_m\}$ of m training signals, where $\boldsymbol{f}_i \in \Re^n$, and a sampling matrix $\boldsymbol{A} \in \Re^{k \times n}$ with k random vectors, and random measurements obtained as:

$$\boldsymbol{M}(j,i) = \sum_{l=1}^{n} \boldsymbol{A}(j,l)\boldsymbol{F}(l,i) \qquad \text{for } j = 1,2,...,k \text{ and } i = 1,2,...,m \tag{4.5}$$

the discrimination power of the measurements produced by each random vector is obtained using the Fisher discriminant ratio (explained in Section 3.1.1) as:

$$F(\boldsymbol{M}) = \frac{trace(\boldsymbol{S}_b)}{trace(\boldsymbol{S}_w)} \tag{4.6}$$

Based on this measurement, the features will be rank ordered and the first p dimensions will be selected and stored in $\tilde{\boldsymbol{M}} \in \Re^{p \times n}$, discarding the $k - p$ dimensions with smaller discrimination values. The indices of the p feature vectors selected will be stored in Ω to be used to restrict the rows of the sampling matrix $\boldsymbol{A}$ and obtain the measurements of the test image and the training images. The test image $\boldsymbol{y}$ will be classified to the l^{th} class if the training image that minimizes:

$$\arg\min_{\boldsymbol{f} \in \boldsymbol{F}} \left\| \boldsymbol{y} - \boldsymbol{A}_{|\Omega} \boldsymbol{f} \right\| \tag{4.7}$$

also belong to that class. As before, the classification accuracy is defined as:

$$Accuracy = \left(1 - \frac{\sum_{i=1}^{z} \boldsymbol{e}_i}{z} \right) \times 100 \tag{4.8}$$

where z is the total number of test images and $\boldsymbol{e}_i = 1$ if the test image is misclassified and $\boldsymbol{e}_i = 0$ otherwise. The mathematical description of this method can be found in Figure 14.

Input: Training signals $\boldsymbol{F} \in \Re^{n \times m}$,Test signal $\boldsymbol{f}_T \in \Re^n$, sampling matrix $\boldsymbol{A} \in \Re^{k \times n}$ Output: Test image label l	
Haupt et al. Algorithm	Proposed Algorithm
	$\boldsymbol{M} = \langle \boldsymbol{A}, \boldsymbol{F} \rangle$ (Training measurements) $\Omega \leftarrow \arg\max_{\boldsymbol{M}} (F(\boldsymbol{M}))$ (Feature Selection)
$\boldsymbol{y} \leftarrow \langle \boldsymbol{A}, \boldsymbol{f}_T \rangle$	$\boldsymbol{y} \leftarrow \langle \boldsymbol{A}_{\|\Omega}, \boldsymbol{f}_T \rangle$ (Test measurements)
$l \leftarrow \arg\min_{\boldsymbol{f} \in \boldsymbol{F}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{f}\|$	$l \leftarrow \arg\min_{\boldsymbol{f} \in \boldsymbol{F}} \|y - \boldsymbol{A}_{\|\Omega} \boldsymbol{f}\|$ (Test labeling)

Figure 14. Haupt et al algorithm (left) and the proposed algorithm (right) for image classification from compressive measurements

4.2 EXPERIMENTS AND RESULTS

In this section, the results obtained with the proposed algorithm will be compared to those of the Haupt et al. algorithm. The database used in this experiment is the ETHZ database explained in Section 3.2.1. In this case, to obtain the compressed sensing measurement, the original images of size 128×128 were used instead of 16×16 nonoverlapping patches. The elements of the sampling matrix were drawn from a Gaussian function with zero mean and unitary standard deviation. The results of this section are given for 100 simulations with a 10-fold cross-validation.

Figure 15 shows the results for (a) 20 measurements, (b) 40 measurements, (c) 60 measurements, (d) 80 measurements, (e) 100 measurements and (f) 256 measurements as the size of the compressed samples from which the most relevant features will be extracted. The solid straight red line in the graphs corresponds to the result of Haupt et al. algorithm used as reference and the blue line with the asterisks corresponds to the results obtained with the proposed feature selection method.

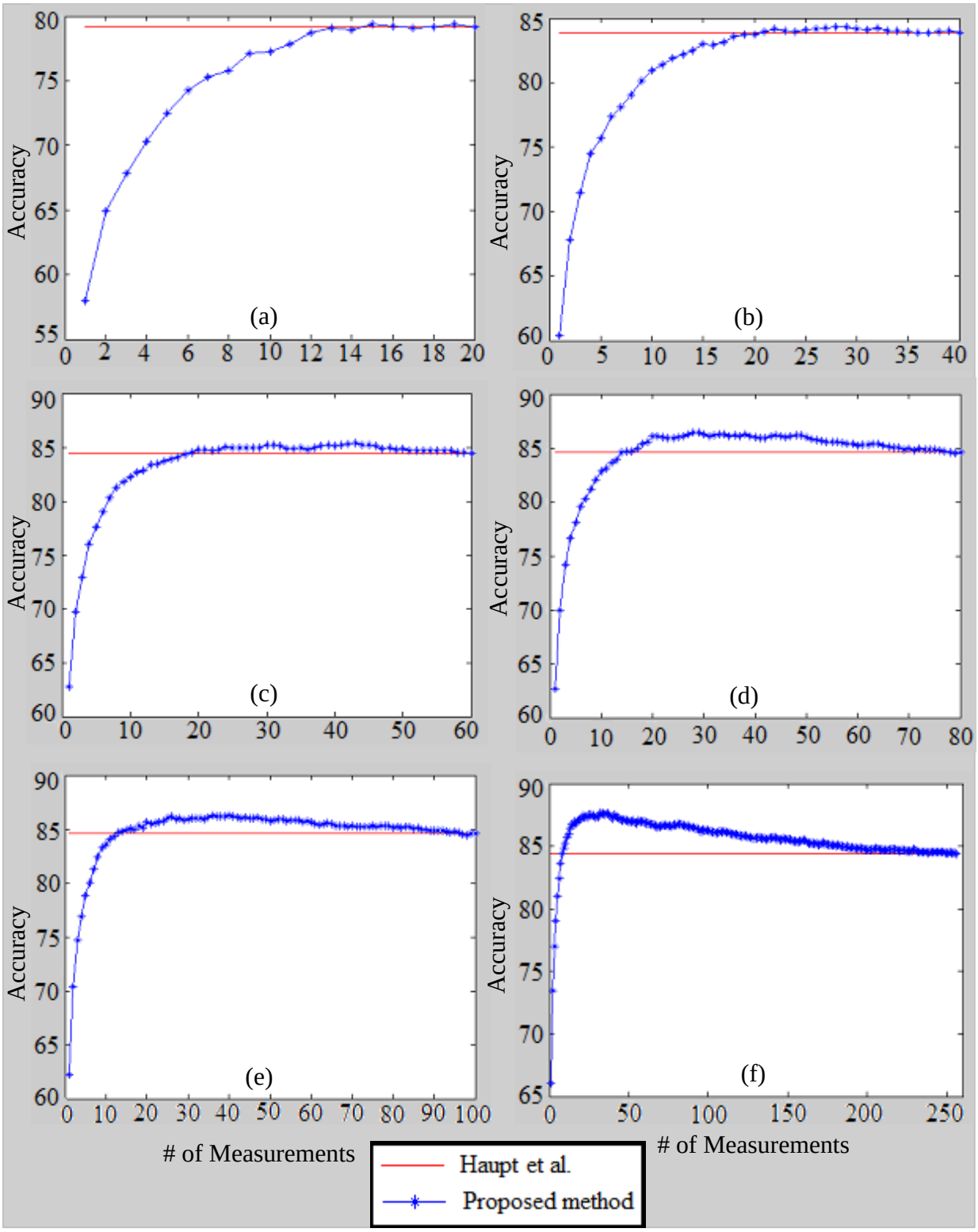

Figure 15. Classification results of the proposed method using different number of measurements: (a) 20 measurements, (b) 40 measurements, (c) 60 measurements, (d) 80 measurements, (e) 100 measurements and (f) 256 measurements.

For all the cases presented in Figure 15, the proposed algorithm obtains similar or better results with less number of measurements/features than using the whole set. When the number of measurements is less than 20, removing some features will degrade the classification accuracy because all the features available are important to be able to discriminate the images. Table 6 presents the classification results obtained using different number of measurements and the size of the minimum subset of these measurements needed to obtain similar or better accuracy with the proposed method. It can be seen that as the number of compressed sensing measurements increases, the amount of reduction in the support of the feature set that also increases. Hence, when more number of features is available, there is more redundancy and a chance to improve classification by adding a feature selection step to compressed sensing.

# of Measurements / Features	Accuracy using the whole set of measurements / features	Smallest # of Measurements/Features to reach the same or better accuracy with the proposed method	Maximum reduction that can be achieved
20 Measurements	79.15%	13	35%
40 Measurements	83.85%	21	47.5%
60 Measurements	84.39%	19	68.33%
80 Measurements	84.61%	15	81.25%
100 Measurements	84.6%	13	87%
256 Measurements	84.4%	9	96.48%

Table 6. Number of measurements needed with the proposed method to achieve better results than using the whole set of measurements

4.3 CONCLUSIONS

In this chapter, the inclusion of a discrimination measure for the selection of features from compressed sensing measurements was proposed. The motivation for this modification to compressed sensing is to remove redundant or irrelevant features from the original compressed features to increase the classification accuracy. The performance of the proposed algorithm was evaluated with noiseless images from the ETHZ database. It was found that using the Fisher discriminant ratio to select a subset of the compressed sensing samples improves the classification accuracy and decreases the sparsity of the feature set. As a result, the selection of a subset of the features produced similar or better results than using the whole set of features. With the proposed algorithm, it was obtained that the number of measurements can be reduced to at most by 35% from 20 measurements and by 96.48% from 256 measurements. In conclusion, it is possible to achieve better classification results using a subset of the features than the whole set.

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

5.1 CONCLUSIONS

This thesis discussed the problem of selecting a compact set of features with high discrimination power for image classification. The methods proposed to solve this problem are sparse representation and compressed sensing, two closely related and well-developed techniques for signal compression and reconstruction. The methods proposed in this thesis extend the framework for sparse signal representation and sensing to image classification applications such that the representations are sparse and discriminative at the same time. Both methods help to reduce the dimensionality of the features to avoid handling high dimensional images for classification purposes. The first method achieves this goal by projecting the sample images against a small number of elements from an overcomplete dictionary that produces a discriminative sparse representation of the data. The second method uses a feature selection measure to obtain, from random compressed sensing measurements, the most relevant measurements (features) such that the classification accuracy is equal or better than using all the measurements.

In the first part of this thesis, a sparse representation method was proposed based on a greedy algorithm similar to CoSaMP to select a group of atoms from an overcomplete dictionary to produce a discriminative sparse representation of a set of images from different classes. This goal was achieved through the inclusion of a cost function that performs a tradeoff between discrimination power and sparsity. First, it was showed that discriminative representations perform better than reconstructive representations for classification purposes. Then, different

experiments were performed using two different image databases with different levels of Gaussian noise and occlusion. The results of these experiments showed that the inclusion of the cost function helps to select atoms that produce discriminative features and the classification accuracy of the proposed method is better than the accuracy obtained from the modified OMP. The proposed algorithm was also compared with LDA using 128×128 images from the ETHZ database. In this case, random patches of size 16×16 were extracted to avoid the construction of a high dimensional dictionary. It was shown that classification accuracies higher than those of LDA can be obtained using low dimensional features extracted from a set of random patches. For the worst case, the dimensionality of the patches was reduced from 256 to 52 which correspond to a reduction of 80%. In conclusion, it was showed that the proposed algorithm can handle images with high intra-class variability and high level of noise and occlusion.

For the second part of this thesis, a feature selection method was proposed to select the most relevant features from the compressed sensing measurements for classification purposes. With the use of a cost function that measures the discrimination power of the set of random measurements, it was shown that a subset of the measurements produces better classification accuracy than the whole set. This is a result of eliminating irrelevant features present in the compressed sensing measurement. The greater the number of compressed sensing measurements available to choose from, the greater the number of irrelevant features that can be eliminated using the feature selection method proposed in Chapter 4.

5.2 FUTURE WORK

As future work, different methods could be evaluated to select the patches instead of using random patches from high dimensional images. The selection of patches with high

discrimination power would help to obtain higher classification accuracies. Some methods that have been used to extract discriminative patches for classification purposes are combinatorial and statistical methods [65] or methods that extract the patches around some interest points detected using operators such as the Förstner operator [66] or SIFT descriptor [67]. Also, future research can include determining the optimal size of the patches such that there is a tradeoff between the dimensionality of the patch and the classification accuracy to avoid computational complexity in the implementation of the algorithm.

For the proposed method presented in Chapter 4, future work could consider different sampling matrices other than the Gaussian random matrix used in this thesis. The sampling matrix to obtain the measurements can be learned from the images such that only a small number of measurements contain higher discrimination information. Previously, it was mentioned that the construction of learned dictionaries can increase the computational complexity of the algorithm. To avoid this problem, only one sampling matrix with low dimensionality can be learned using a subset of patches instead of the whole images.

REFERENCES

REFERENCES

[1] I.T. Jolliffe, *Principal Component Analysis*, Series: Springer Series in Statistics, 2nd ed., Springer, NY, 2002.

[2] M. Vlachos, C. Domeniconi, D. Gunopulos, G. Kollios, and N. Koudas, "Non-Linear Dimensionality Techniques for Classification and Visualization," *Proceeding of the Eight ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 645-651, 2002.

[3] B. Schoelkopf, A. Smola, and K. R. Mueller, "Nonlinear Component Analysis as a Kernel Eigenvalue Problem," *Neural Computation*, vol. 10, pp. 1299-1319, July 1998.

[4] W. Pan, T.D. Bui, and C. Y. Suen, "Text Detection from Scene Images using Sparse Representation," *19th International Conference on Pattern Recognition*, pp. 1-5, December 2008.

[5] M. Zhao and S. Li, "Sparse Representation Classification for Text Detection," *Second International Symposium on Computational Intelligence and Design*, vol.1, pp. 76-79, December 2009.

[6] G. K. Vinay, S. M. Haque, R. V. Babu, and K. R. Ramakrishnan, "Human Detection using Sparse Representation," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, March 2012.

[7] V. Cevher, A. Sankaranarayanan, M. F. Duarte, D. Reddy, R. G. Baraniuk, and R. Chellappa, "Compressive Sensing for Background Subtraction," *Lecture Notes in Computer Science*, vol. 5303, pp. 155-168, 2008.

[8] S. Mallat, *A Wavelet Tour of Signal Processing: The Sparse Way*, Academic Press an imprint of Elsevier, 3rd ed., Burlington, MA, 2009.

[9] B. Vidakovic and P. Mueller, "Wavelets for Kids: A Tutorial Introduction," Institute of Statistics and Decision Science, Duke University, pp. 1-28, 1994.

[10] S. M. Joseph, R. Sebastian, and B. Anto, "The Effect of Different Wavelets on Speech Compression," *ACM Proceedings of the 2011 International Conference on Communication, Computing & Security*, pp.265-268, 2011.

[11] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, "Image Coding Using Wavelet Transform, " *IEEE Transactions on Image Processing*, vol. 1, April 1992.

[12] A. S. Lewis, and G. Knowles, "Image compression using the 2-D Wavelet Transform," *IEEE Transactions on Image Processing*, vol. 1, April 1992.

[13] J. Reichel, G. Menegaz, M. J. Nadenau, and M. Kunt, "Integer Wavelet Transform for Embedded Lossy to Lossless Image Compression," *IEEE Transactions on Image Processing*, vol. 10, pp. 383-392, March 2001.

[14] X. G. Xia, C. Boncelet, and G. Arce, "Wavelet Transform based watermark for digital images," *Optic Express*, vol. 3, pp. 497-511, December 1998.

[15] P. Bao, and X. Ma, "Image Adaptive Watermarking Using Wavelet Domain Singular Value Decomposition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, pp. 96-102, January 2005.

[16] J. L. Starck, E. J. Candes, and D. L Donoho, "The Curvelet Transform for Image Denoising," *IEEE Transactions on Image Processing*, vol. 11, pp. 670-684, June 2002.

[17] K. Huang, Z. Wu, G. S. K. Fung, and F. H. Y. Chang, "Color image denoising with wavelet thresholding based on human visual system model," *Signal Processing: Image Communication*, vol. 20, pp. 115-127, February 2005.

[18] A. Apatean, A. Rogozan, S. Emerich, and A. Bensrhair, "Wavelets as features for objects recognition," *Acta Tehnica Napocensis*, vol. 49, pp. 23-36, February 2008.

[19] S. Rastegar, A. Babaeian, M. Bandarabadi, and Y. Toopchi, "Airplane Detection and Tracking Using Wavelet Features and SVM Classifier", *41st Southeastern Symposium on System Theory*, pp 64-67, March 2009.

[20] K. M. Rajpoot, and N. M. Rajpoot, "Wavelets and Support Vector Machines for Texture Classification," *In 8th IEEE International Multioptic Conference (INMIC'04)*, pp. 328–333, 2004.

[21] S. Arivazhagan, and L. Ganesan, "Texture Classification using Wavelet Transform," *Pattern Recognition Letters*, vol. 24, pp. 1513-1521, June 2003.

[22] K. Huang and S. Aviyente, "Wavelet Feature Selection for Image Classification," *IEEE Transactions on Image Processing*, vol. 17, pp. 1709-1720, September 2008.

[23] X. Wang, "Image Edge Detection Based on Lifting Wavelet," *Intelligent Human-Machine Systems and Cybernetics*, vol. 1, pp. 25-27, August 2009.

[24] M. Rizzi, M. D'Aloia, and B. Castagnolo, "Computer Aided Detection of Microcalcifications in Digital Mammograms Adopting a Wavelet Decomposition," *Integrated Computer-Aided Engineering*, vol. 16, pp. 91-103, 2009.

[25] M. N. Do, and M. Vetterli, "The Contourlet Transform: An Efficient Directional Multiresolution Image Representation," *IEEE Transactions on Image Processing*, vol. 14, pp. 2091-2106, December 2005.

[26] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An Algorithm for Designing of Overcomplete Dictionaries for Sparse Representation," *IEEE Transactions on Signal Processing*, vol. 54, pp. 4311-4322, November 2006.

[27] S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, pp. 33–61, 1998.

[28] D. Wipf and B. Rao, "Sparse bayesian learning for basis selection," *IEEE Transactions on Signal Processing*, vol. 52, pp. 2153–2164, August 2004.

[29] J. Shi, X. Ren, G. Dai, J. Wang, and Z. Zhang, "A non-convex relaxation approach to sparse dictionary learning," *In Proceedings of the IEEE international conference on computer vision*, pp. 1809–1816, June 2011.

[30] S. Mallat, and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, pp. 3397–3415, December 1993.

[31] Y. C. Pati, R. Reiizafar, and P. S. Krishnaprasad, "Orthogonal Matching Pursuit: Recursive Function Approximation with Applications to Wavelet Decomposition," *Asilomar Conference on Signals Systems and Computers*, pp. 1-5, November 1993.

[32] D. Needell, and J. Tropp, "Cosamp: Iterative signal recovery from incomplete and inaccurate samples," *Applied and Computational Harmonic Analysis*, vol. 26, pp. 301–321, May 2009.

[33] M. A. Hameed, "Comparative Analysis of Orthogonal Matching Pursuit and Least Angle Regression," *MS Thesis*, Michigan State University, 2012.

[34] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society*, vol. 58, pp. 267-288, 1996.

[35] D.L. Donoho, "Compressed Sensing," *IEEE Transactions on Information Theory*, vol. 56, pp. 1289-1306, April 2006.

[36] E. J. Candès, "The restricted isometry property and its applications for compressed sensing," *Comptes Rendus Mathematique*, vol. 346, pp. 589-592, May 2008.

[37] M. A. Davenport, M. F. Duarte, Y. C. Eldar, and G. Kutyniok, "Introduction to Compressed Sensing," Book Chapter in Compressed Sensing: Theory and Applications, Edited by Y. C. Eldar and G. Kutyniok, Cambridge University Press, 2011.

[38] M. A. Davenport, P. T. Boufounos, M. B. Wakin, and R. G. Baraniuk, "Signal Processing with Compressive Measurements," *IEEE Journal of Selected Topics on Signal Processing*, vol. 4, pp. 445-460, April 2010.

[39] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Constructive Approximation*, vol. 28, pp. 253-263, 2008.

[40] E. J. Candes and M. B. Waking, "An Introduction to Compressive Sampling," *IEEE Signal Processing Magazine*, vol. 25, pp. 21-30, March 2008.

[41] M. Lusting, D. L. Donoho, J. M. Santos, and J. M. Pauly, "Compressed Sensing MRI," *IEEE Signal Processing Magazine*, vol. 25, pp. 72-82, March 2008.

[42] R. Baraniuk and P. Steeghs, "Compressive Radar Imaging," *2007 IEEE Radar Conference*, pp. 128-133, April 2007.

[43] M. F. Duarte, M. A, Davenport, D. Takhar, J. N. Laska, T. Sun, K. F. Kelly, and R. G. Baraniuk, "Single-Pixel Imaging via Compressive Sampling," *IEEE Signal Processing Magazine*, vol. 25, pp. 83-91, March 2008.

[44] M. Elad and M. Aharon, "Image denoising via sparse and redundant representation over learned dictionaries," *IEEE Transactions on Image Processing*, vol. 15, pp. 3736-3745, December 2006.

[45] M. J. Fadili, J.-L. Starck, and F. Murtagh, "Inpainting and Zooming using Sparse Representations," *The Computer Journal*, vol. 52, pp. 64-79, 2009.

[46] J. Xu, Y. Pi, and R. Ming, "SAR Image Compression Based on Sparse Representation," *11th International Radar Symposium*, pp. 1-4, June 2010.

[47] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust Face Recognition via Sparse Representation," *IEEE Transactions* on *Pattern Analysis and Machine Intelligence*, vol. 30, pp. 210–227, February 2009.

[48] K. Huang and S. Aviyente, "Sparse Representation for Signal Classification," *In Advances in Neural Information Processing Systems*, pp. 609–616, 2006.

[49] F. Rodriguez and G. Sapiro, "Sparse Representations for Image Classification: Learning Discriminative and Reconstructive Nonparametric Dictionaries," *Technical report*, University of Minnesota, December 2007.

[50] M. Strauss, J. Tropp, and A. Gilbert, "Algorithms for simultaneous sparse approximation part I: Greedy pursuit," *Signal Processing*, vol. 86, pp. 572–588, April 2006.

[51] M. Strauss, J. Tropp, and A. Gilbert, "Algorithms for simultaneous sparse approximation part II: Convex Relaxation," *Signal Processing*, vol. 86, pp. 589–602, April 2006.

[52] S. Cardona-Romero and S. Aviyente, "Discriminative Sparse Image Representation for Classification Based on a Greedy Algorithm," *IEEE Statistical Signal Processing Workshop*, August 2012.

[53] S. A. Nene, S. K. Nayar, and H. Murase, "Columbia Object Image Library (COIL-20)," *Technical Report CUCS-005-96*, February 1996. Database available at http://www1.cs.columbia.edu/CAVE/software/softlib/coil-20.php.

[54] B. Leibe and B. Schiele, The Pascal Object Recognition Database Collection: The TU Darmstadt Database (formerly the ETHZ Database), *Database*. Available at http://pascallin.ecs.soton.ac.uk/challenges/VOC/databases.html#TUD.

[55] H. Yi, "Robust Wavelet Transform-based Correlation Edge Detectors using Correlation of Wavelet Coefficients," *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 4, pp. 77-88, December 2011.

[56] P. L. Dragotti and M. Vetterli, "Wavelets Footprints: Theory, Algorithms, and Applications," *IEEE Transactions on Signal Processing*, vol. 51, pp. 1306-1323, May 2003.

[57] V. Kumar, "Face Recognition using Gabor Wavelets," *Technical Report*, Global Academy of Technology, 2006.

[58] C.-C. Chang and C.-J. Lin, "LIBSVM -- A Library for Support Vector Machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1-27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[59] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A Practical Guide to Support Vector Classification," *Technical Report*, National Taiwan University, 2010. Available at http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

[60] M. Lustig, D. Donoho, and J. M. Pauly,"Sparse MRI: The Applications of Compressed Sensing for Rapid MR Imaging," *Magnetic Resonance in Medicine*, vol. 58, pp. 1182-1195, December 2007.

[61] L. Liu and P. W. Fieguth, "Texture Classification from Random Features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, pp. 574-586, March 2012.

[62] M. F. Duarte, M. A. Davenport, M. B. Wakin, and R. G. Baraniuk, "Sparse Signal Detection from Incoherent Projections," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toulouse, France, May 2006.

[63] J. Haupt, R. Castro, R. Nowak, G. Fudge, and A. Yeh, "Compressive Sampling for Signal Classification," In *Proc. 40th Asilomar Conf. Signals, Systems and Computers*, November 2006.

[64] H. Almuallim and T.G. Dietterich, "Learning with many irrelevant features", In: *Ninth National Conference on Artificial Intelligence*, pp. 547-552, 1991.

[65] A. Vashist, Z. Zhao, A. Elgammal, I. Muchnik, and C. Kulikowski, "Discriminative Patch Selection using Combinatorial and Statistical Models for Patch-Based Object Recognition," *Conference on Computer Vision and Pattern Recognition Workshop*, pp. 1-12, June 2006.

[66] S. Agarwal and D. Roth, "Learning a Sparse Representation for Object Detection", *Proceedings of the Seventh European Conference on Computer Vision*, vol. 4, pp. 113-130, 2002.

[67] F. Xu and Y.-J. Zhang, "Feature Selection for Image Categorization", *Computer Vision - ACCV 2006*, vol. 3852, pp. 653-662, 2006.

[68] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 711-720, July 1997.

[69] A. M. Martinez and A. C. Kak, "PCA versus LDA." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 228-233, February 2001.