Joint Sparsity-based Classification of Color Images

Charles Jeon,* Vishal Monga,† and Umamahesh Srinivas‡

July 22, 2012

Abstract

For many years, sampling theory has been based on Shannon and Nyquist who stated that band-limited signals can be exactly reconstructed using samples acquired at or higher than Nyquist rate (1949). Recently however, the focus has shifted to compressed sensing, where if the underlying signal is sparse, the signal can be represented by a small collection of linear projections. It is now understood that many type of signals, such as images and audio signals, fall into this category. Compressed sensing techniques are very useful in applications where taking a large number of measurements of the object is very costly, such as in medical imaging.

Recently norm-minimization ideas from compressed sensing have been adapted for classification tasks via class-specific image dictionaries. The key assumption is that the vectorized versions of training images belonging to the same class lie in a low-dimensional subspace and therefore, any new representation from this image class can be approximated by a linear combination of the training images. The sparse representation of these are naturally discriminative and so, classification can be performed based on reconstruction error. The seminal contribution by Wright et al. to face recognition has demonstrated the robustness of this approach to a variety of real-world distortions. In many applications, prior knowledge about a collection of test images belonging to the same class can be exploited to improve classification performance in a joint sparsity setting. In this work, we exploit inter-channel information between the red, green, and blue channels of color images in a joint sparsity framework for the classification of histopathological images.

1 Introduction

One of the key theorems that brought many developments in signal processing is the Shannon and Nyquist Sampling Theorem: a band-limited signal can be exactly and uniquely reconstructed by taking samples at twice the highest frequency [1]. For many years, Sampling Theorem has been essential in almost all of the signal acquisition methods used in various electronic applications. For example, a standard digital-to-analog (DAC) and analog-to-digital converter (ADC) relies on the Sampling Theorem extensively to reconstruct or quantize the given signal. However, in the last few years, an alternative theory has been introduced that did not follow the conventional protocol for data acquisition. Compressive Sensing (CS), first introduced by Donoho [2], shows that in special cases, signals can be recovered by far fewer samples or measurements than...
using traditional methods [3]. The key underlying principle that enables this is that most signals have few non-zero coefficients with respect to some inherent basis, which in CS terminology, is the concept of sparsity.

Sparsity suggests that “information rate” of the underlying signal may be much smaller than its bandwidth. In other words, the signal is compressible, such that the degree of freedom is smaller than the length of the signal. For instance, a sparse signal can be expressed as a small number of linear combinations in some fixed basis. One of the most common examples of these is found in imaging, using DCT basis for JPEG and wavelets for JPEG2000 [4]. The crucial impact of sparsity is that one can design a new sampling method that can capture the information in a sparse signal and condense it to a small amount of data.

The applications of sparse data representation have been extended to the area of pattern recognition with the development of compressed sensing framework and image processing for sparse modeling of signals. These applications are based on the observation that while the dimension of each individual signal is high, signals in the same class usually share a low-dimensional subspace [5]. Since CS techniques ensure the recoverability of these signals from linear projections, the sample signal can be decomposed over an overcomplete dictionary generated by a set of representative samples. Many applications have been shown in literature in various fields, such as image restoration [6], super-resolution [7], target detection [8], and face recognition [9].

Recently, the applications of sparsity framework have been extended to classification tasks. Wright et al., have demonstrated that sparse signal representations have a robust performance in face recognition. Moreover, the Sparse Representation-based Classification (SRC) approach has been shown to outperform the conventional methods, Principal Component Analysis (PCA) and Independent Component Analysis (ICA) methods under random corruption and contiguous occlusion in images [9]. Face recognition is often times a hard problem due to high variability in the process of acquiring face images, which necessitates robust algorithms. Suppose that we are to collect a number of images for an individual in various settings such as, luminosity, color balance and occlusions. Provided that we have a sufficient number of images, it will be possible to represent any test image of the individual as a linear combination of the previously collected images. In macroscopic view with many individuals, this representation is sparse, involving only a small number of samples of our image database. In many cases, seeking the sparsest representation automatically discriminates between the various classes present in the training set [9].

In hyperspectral image classification, it is quite typical to observe that a local neighborhood of objects observe similar class behavior. Conversely, if some objects are known to belong to an identical class or share a common class behavior, the additional information can be exploited in using sparse models in classification. Joint sparsity comes from the fact that additional information can be obtained by exploiting correlation, e.g. a common sparsity pattern, across neighboring objects [10].

By using joint sparsity model, we propose the idea of extending image classification to multispectral
(color) images. While much research has been focused on grayscale images for representing their spatial structure, additional information could be exploited by utilizing color images. One of the difficulties in straightforward adoption of grayscale classification methods to each individual color channel is that it ignores inherent correlation among the color channels. Also, by considering multiple channels, the overall complexity increases. Although exploiting multispectral images for face recognition has already been introduced in literature [11], we propose a new joint-sparse model for color framework. Since SRC approach has shown robust performance in grayscale images, we expect to observe that color-based SRC will also demonstrate a strong performance in classification.

In this project, we extend the previous ideas of a sparsity based model to construct a joint sparsity model using color channels for image classification. In Section 2, basic classification model along with jointly sparse model proposed in literature [9], [10] are reviewed. In the next section, Section 3, color-based sparse model is formulated along with an additional optimization problem to guarantee optimality. In Section 4, we propose our methodology to solve the introduced problem along with simulation results of histopathological images for bovine organs.

While our simulation results are applied to medical images, our framework is of broad interest to object classification in general. Many feature extraction methods and the use of contrast agents with machine-learning mechanisms have been used in medical imaging [12], [13], but we concentrate on a general SRC based approach here.

2 Classification Based on Sparse Representation

Classification, a general form of automatic (machine) recognition, description, and pattern recognition, is an example of supervised learning, where a training set of correctly identified observations is available, to identify which of a set of categories an observation belongs [14]. Although unsupervised classification methods exist, our primary focus is on supervised classification. Formally speaking, a classification problem can be defined as using some set of training samples from $M$ distinct object classes to correctly determine the class to which the new test sample(s) belongs. If we denote $N_k$ as the number of training samples from the $k$-th class, we can arrange the training samples $\{a^k_i\}, i = 1, \ldots, N_k$ as the columns of a matrix, $A^k = [a^k_1 a^k_2 \cdots a^k_{N_k}]$. We will refer to $A^k$ as our dictionary matrix. We follow lexicographical ordering for characterizing images in the following subsections.
2.1 General Sparsity Classification Model

Many previous works have been done to exploit the structure of $A^k$ for classification. One particular approach we consider here is approximately modeling the samples from a single class as lying on a linear subspace as subspace models are flexible enough to capture much of the variation in real data [5]. Thus, given sufficient training samples of the $m$-th class, $A^m \in \mathbb{R}^{B \times N_m}$, any new sample $y \in \mathbb{R}^B$ from the same class can be approximately expressed as a linear combination of the training samples,

$$ y = \alpha_1^m a_1^m + \alpha_2^m a_2^m + \cdots + \alpha_{N_m}^m a_{N_m}^m $$

$$ = \begin{bmatrix} a_1^m & a_2^m & \cdots & a_{N_m}^m \end{bmatrix} \begin{bmatrix} \alpha_1^m \\ \alpha_2^m \\ \vdots \\ \alpha_{N_m}^m \end{bmatrix}^T = A^m \alpha^m $$

for some weight vector $\alpha^m \in \mathbb{R}^{N_m}$.

Although the class of the test sample is initially unknown, we can deduce that it is modeled to lie in the union of the $M$ distinct subspaces associated with the $M$ classes. Therefore, we can combine the sub-dictionary matrices $A^m$, $m \in \{1, \ldots, M\}$ to define a new dictionary matrix $A$ for the entire collection of training samples:

$$ y = A^1 \alpha^1 + A^2 \alpha^2 + \cdots + A^m \alpha^m $$

$$ = \begin{bmatrix} A^1 & \cdots & A^m \end{bmatrix} \begin{bmatrix} \alpha^1 \\ \vdots \\ \alpha^m \end{bmatrix} = A\alpha $$

(1)

where $A \in \mathbb{R}^{B \times N}$ is a dictionary matrix composed of all the training samples with $N = \sum_{i=1}^{M} N_m$ and weight vector $\alpha \in \mathbb{R}^N$ formed by stacking the individual $\alpha^j$ vectors for all $j \in \{1, \ldots, M\}$. Note that ideally, if $y$ belongs to $j$-th class, then $\alpha^i = 0$ for $i = 1, \ldots, M, i \neq j$. In our sparsity model, $\alpha$ is a sparse vector, which corresponds to few non-zero entries. Note that Equation (1) can also be written as a linear combination of only the $K$ active dictionary elements or atoms, $\{a_{\lambda_k}\}$ corresponding to the $K$ non-zero entries of $\{a_{\lambda_k}\}$, $k \in \{1, \ldots, K\}$

$$ y = \alpha_{\lambda_1} a_{\lambda_1} + \alpha_{\lambda_2} a_{\lambda_2} + \cdots + \alpha_{\lambda_K} a_{\lambda_K} $$

$$ = \begin{bmatrix} a_{\lambda_1} & \cdots & a_{\lambda_K} \end{bmatrix} \begin{bmatrix} \alpha_{\lambda_1} \\ \vdots \\ \alpha_{\lambda_K} \end{bmatrix} = A_{\lambda_K} \alpha_{\lambda_K} $$

(2)
where $K$ can be expressed as $K = \|\alpha\|_0$ which denotes the sparsity level or $l_0$ norm of $\alpha$. The index set $\Lambda_K = \{\lambda_1, \ldots, \lambda_K\}$ denotes the support of $\alpha$, $A_{\Lambda_K}$ is a $B \times K$ matrix whose columns are the $K$ dictionary elements of $\{a_k\}$, $k \in \Lambda_K$ and $\alpha_{\Lambda_K}$ is a $K$-dimensional vector consisting of the entries of $\alpha$ indexed by $\Lambda_K$.

### 2.1.1 Classification model for Joint Sparsity

Recall that we can exploit the spatial correlation across neighboring pixels if they share a common sparsity pattern [10]. In other words, we would be able to make a better guess if we are given groups of test samples that are known to belong to the same class.

Suppose that $y_i$ and $y_j$ are images that belong to the same class. Recall from Equation (2) that $y_i$ can be expressed as $y_i = A\alpha_i = \alpha_{i,\lambda_1}a_{\lambda_1} + \alpha_{i,\lambda_2}a_{\lambda_2} + \cdots + \alpha_{i,\lambda_K}a_{\lambda_K}$ for some index set $\Lambda_K = \{\lambda_i\}_i$, some sparse coefficient vector $\alpha_i$, and some dictionary $A$. Since $y_j$ was defined to be in the same class as $y_i$, $y_j$ can also be approximated by the same set of training samples $\{a_{\Lambda_k}\}_k$:

$$y_j = A\alpha_j = \alpha_{j,\lambda_1}a_{\lambda_1} + \alpha_{j,\lambda_2}a_{\lambda_2} + \cdots + \alpha_{j,\lambda_K}a_{\lambda_K}$$

Now, we extend this idea to a general joint sparsity model. Suppose we have $T$ different images given by $Y = [y_1 \cdots y_T]$ a $B \times T$ matrix, where each column $y_j$ denotes the $j$-th image belonging to some class $m$. Then, using the ideas explained previously, we have that

$$Y = \begin{bmatrix} y_1 & y_2 & \cdots & y_T \end{bmatrix} = \begin{bmatrix} A\alpha_1 & A\alpha_2 & \cdots & A\alpha_T \end{bmatrix} = A[\alpha_1 \alpha_2 \cdots \alpha_T] = AS$$

where $S = [\alpha_1 \cdots \alpha_T]$ is a $N \times T$ matrix consisting of the sparse vectors $\{\alpha_j\}$ corresponding to each image $y_j$. Note that the sparse vectors $\{\alpha_j\}$ share the same support $\Lambda_K$, thus $S$ is a structured sparse matrix with only $K$ non-zero rows.

### 2.2 Reconstruction

The first step in the classification process is to consider a reconstruction problem of finding the sparse matrix $S$ for some set of test sample $Y$. Given a dictionary of training image samples $A$ for the test samples $y_i$, we wish to solve the following optimization problem,

$$\hat{S} = \arg \min \|S\|_{row,0} \quad \text{subject to} \quad AS = Y$$

(3)
where the operator $\|S\|_{row,0}$ denotes the number of non-zero rows of $S$. Note that for $T = 1$, where we have a single image, the problem reduces to a general sparsity problem,

$$\hat{\alpha} = \arg \min \|\alpha\|_0 \quad \text{subject to} \quad A\alpha = y$$

Therefore, we will consider the general joint-sparsity based model throughout the paper. Now, we can modify Equation (3) to relax the equality constraint for approximation errors in empirical data to

$$\hat{S} = \arg \min \|S\|_{row,0} \quad \text{subject to} \quad \|AS - Y\|_F \leq \sigma \quad (4)$$

where $\sigma > 0$ is the error tolerance and $\|\cdot\|_F$ denotes the Frobenius norm. This can be also interpreted as minimizing the approximation error given some sparsity level [15],

$$\hat{S} = \arg \min \|AS - Y\|_F \quad \text{subject to} \quad \|S\|_{row,0} \leq K_0 \quad (5)$$

where $K_0$ is an upper bound on the sparsity level. While the problems in Equations (3), (4), and (5) are all $NP$-hard, many approximations have been introduced to make these problems solvable in polynomial time [16], [17], [18]. One particular approximation method that we will exploit in this paper is the greedy pursuit algorithm known as Simultaneous Orthogonal Matching Pursuit (S-OMP) [19], [20]. The intuition behind S-OMP is that it picks the atoms that contribute the most energy to every column of the signal matrix in each iteration. The S-OMP algorithm finds the support of the sparse vector that approximately solves Equation (5). At each iteration, S-OMP picks the dictionary element that maximizes the residual norm and creates an orthogonal projection onto the residual matrix. After $K_0$ iterations, S-OMP returns the set of $K_0$ atoms that contribute most of the energy in the signal. The algorithm is summarized below in Algorithm 1.

Note for $l_p$ norm, $p = 1$ is used in [19], but we use $p = 2$, used in [21].

Now, the obtained row-sparse matrix $\hat{S}$ can be used to determine the class of $Y$ by considering the error residuals between the original test samples and the approximation obtained from each class sub-dictionary:

$$r^m(Y) = \|Y - A^m\hat{S}^m\|_F, \quad m = 1, 2, \ldots, M \quad (6)$$

where $\hat{S}^m$ consists of the $N_m$ rows in $\hat{S}$ that are associated with the $m$-th class sub-dictionary $A^m$ in Section 2. After computing all the $\{r^m(Y)\}_m$, the label of $Y$ is given as class with the minimum total residual,

$$\text{Class}(Y) = \arg \min_{i=1,\ldots,M} r^i(Y) \quad (7)$$
**Algorithm 1**: S-OMP Algorithm

**Input**: \( B \times N \) dictionary matrix \( A = [a_1 \cdots a_N] \), \( B \times T \) signal matrix \( Y = [y_1 \cdots y_T] \), and number of iterations \( K \)

**Initialization**: residual \( R_0 = Y \), index set \( \Lambda_0 = \phi \), iteration counter \( k = 1 \)

while \( k \leq K \) do

(1) Find the index of the atom that best approximates all residuals: \( \lambda_k = \arg \max_{i=1,\ldots,N} \| R_{k-1}^T a_i \|_p, p \geq 1 \)

(2) Update the index set \( \Lambda_k = \Lambda_{k-1} \cup \{ \lambda_k \} \)

(3) Compute the orthogonal projector \( P_k = (A_{\Lambda_k}^T A_{\Lambda_k})^{-1} A_{\Lambda_k}^T Y \in \mathbb{R}^{k \times T} \) where \( A_{\Lambda_k} \in \mathbb{R}^{B \times k} \) consists of the \( k \) atoms in \( A \) indexed in \( \Lambda_k \)

(4) Update the Residual Matrix \( R_k = Y - A_{\Lambda_k} P_k \)

(5) Increment \( k \): \( k \leftarrow k + 1 \)

end while

**Output**: Index set \( \Lambda = \Lambda_K \), the sparse representation \( \hat{S} \) whose non-zero rows index by \( \Lambda \) are the \( K \) rows of the matrix \( (A_{\Lambda_K}^T A_{\Lambda_K})^{-1} A_{\Lambda_K}^T Y \)

3. **Color-based Sparse Representation**

Every color image \( X \) is composed of three distinct channels, \( X = [x_r \ x_g \ x_b] \in \mathbb{R}^{B \times 3} \) where the subscripts \( r, g, b \) correspond to the red, green, and blue color channels respectively. The training dictionary matrix \( A \) is redefined as \( A = [A_r \ A_g \ A_b] \in \mathbb{R}^{B \times 3N} \). For simplicity, let us assume that we have \( L \) training samples from each of the \( M \) distinct classes so that \( N = ML \). Therefore, we can define each color dictionary: \( A_c = [A_{c1} \ A_{c2}^r \ldots A_{cM}^r] \in \mathbb{R}^{B \times ML} \), \( c \in \{r, g, b\} \) as the concatenation of the sub-dictionaries from all classes belonging to the same color channel. We note that the color dictionaries are designed to obey column correspondence, i.e., the \( i \)-th column of the training samples for the three color channels correspond to the \( i \)-th image.

3.1 **Joint Sparsity Model for Color Images**

With the notations in place, the test color image \( X \) can now be represented as a linear combination of training samples as follows:

\[
X = AS = [A_r^1 \cdots A_r^M A_g^1 \cdots A_g^M A_b^1 \cdots A_b^M] [\alpha_r \ \alpha_g \ \alpha_b]
\]

where the coefficient vectors \( \alpha_r, \alpha_g, \alpha_b \in \mathbb{R}^{3ML} \) and \( S = [\alpha_r \ \alpha_g \ \alpha_b] \in \mathbb{R}^{3ML \times 3} \).

We first start by examining the structure of the coefficient matrix \( S \). As we can assume that each color channel \( c \in \{r, g, b\} \) of the test color image can be approximately represented by the span of the training
samples belonging to the same color channel, the columns of $S$ has the following structure,

$$\alpha_r = \begin{bmatrix} \alpha^1_r \\ \vdots \\ \alpha^M_r \end{bmatrix}, \quad \alpha_g = \begin{bmatrix} 0 \\ \vdots \\ \alpha^M_g \end{bmatrix}, \quad \alpha_b = \begin{bmatrix} 0 \\ \vdots \\ \alpha^M_b \end{bmatrix}$$

where each of the sub-vectors $\{\alpha^k_c\}_{k} \in \mathbb{R}^L$ for all $c \in \{r, g, b\}$ and $0 \in \mathbb{R}^{ML}$ denotes the zero vector. We note that $S$ exhibits block-diagonal structure.

Recalling the similarity to our model formation for grayscale images in Section 2, each color channel of the test image can be represented by a sparse linear combination of the sub-dictionaries of the training samples in that color channel. Furthermore, the non-zero weights of color training samples in the linear combination exhibit one-to-one correspondence across channels. In mathematical notation, this implies that if $l$-th training sample from the $m$-th class has a non-zero coefficient in its weights for some color channel $c$, it is also necessarily non-zero for the other color channels. This suggests a joint sparsity model similar to the model introduced in Section 2.1.1. However, the construction of $S$ does not permit us to apply previous $l_0$ row-sparsity framework. Since $S$ obeys column correspondence, we introduce matrix $S' \in \mathbb{R}^{ML \times 3}$ as the transformation of matrix $S$ with the zero coefficients removed,

$$S' = \begin{bmatrix} \alpha^1_r & \alpha^1_g & \alpha^1_b \\ \vdots & \vdots & \vdots \\ \alpha^M_r & \alpha^M_g & \alpha^M_b \end{bmatrix}$$

With the clever formation of $S'$, we can now apply row-sparsity similar to Equation (4). By taking into account the approximation error, our problem becomes

$$\tilde{S} = \arg \min \|S'\|_{row,0} \quad \text{subject to} \quad \|X - AS\|_F \leq \epsilon \quad (9)$$

for some tolerance $\epsilon > 0$. Note that we are trying to minimize the number of non-zero rows, while the constraint guarantees a good approximation.
3.1.1 Transformation of \( S \)

We wish to address the particular method of transforming \( S \) by introducing matrices \( H \in \mathbb{R}^{3ML \times 3} \) and \( J \in \mathbb{R}^{ML \times 3ML} \),

\[
H = \begin{bmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1 \\
\end{bmatrix},
J = \begin{bmatrix} I_{ML} & I_{ML} & I_{ML} \end{bmatrix}
\]

where \( 1, 0 \in \mathbb{R}^{ML} \) are vector containing all ones and zeros respectively, and \( I_{ML} \) denotes the \( ML \)-dimensional identity matrix. Thus we have the operation \( S' = J (H \circ S) \) where \( \circ \) denotes the Hadamard product, where \( (H \circ S)_{ij} \triangleq h_{ij}s_{ij} \) for all \( i, j \).

3.2 Refining the Choice of \( S \)

The cost function of Equation (9) is non-convex, and thus any solution obtained is suboptimal. To improve the solution, we propose a convex auxiliary optimization problem, thus guaranteeing a unique minimum solution for the additional problem. From the initially obtained solution \( \hat{S}' \) we construct \( \hat{S} \) by inserting zeros appropriately. Then, we design a membership matrix \( E \in \mathbb{R}^{3ML \times 3} \) which has zeros at locations of non-zero entries in \( \hat{S} \) and ones elsewhere, that is, \( e_{ij} = 1 \{ \hat{s}_{ij} = 0 \} \) for all \( i, j \), where \( 1 \) is the indicator function. With the introduction of the membership matrix \( E \), we have the following optimization problem,

\[
\hat{S} = \arg \min \| X - AS \|_F \quad \text{subject to} \quad s_i^T e_i = 0, i = 1, 2, 3
\]  

where \( s_i \) and \( e_i \) denotes the \( i \)-th column of \( S \) and \( E \).

To mitigate computational complexity, we can simplify the problem further. Each column of \( S \) can be optimized in parallel since the constraints are separable. Therefore, the preceding optimization problem can be simplified to:

\[
\hat{s}_1 = \arg \min \| x_r - As_1 \|_2 \quad \text{subject to} \quad e_1^T s_1 = 0
\]

and similarly for \( \hat{s}_2 \) and \( \hat{s}_3 \) for green and blue channels respectively. If we denote the three columns of \( S' \) as \( s_r, s_g, s_b \) and the corresponding columns of \( E \) as \( e_r, e_g, e_b \), we can exploit the knowledge of the locations of the non-zero coefficients to remove the redundant columns from dictionary matrix \( A \) and solve three
quadratic programming problems for \( c \in \{r, g, b\} \) in parallel:

\[
\hat{s}_c = \arg \min \| x_c - A_c s_c \|_2 \quad \text{subject to} \quad e_c^T s_c = 0
\]

(12)

3.3 S-OMP Algorithm for Color Jointly Sparse Model

While Equation (9) looks quite similar to our proposed joint sparsity model in Equation (4), the Hadamard operator from \( S \) to \( S' \) makes the problem much more complex. Without the Hadamard operator in Equation (12), we can relax the row-sparse \( l_0 \) norm to a general \( \| \cdot \|_{p,q} \) norm [22]. In an attempt to solve the problem, we modify the S-OMP Algorithm discussed in Section 2.2. We note the fact that we can make a unique \( S \) from \( S' \) by inserting zeros appropriately, while the converse does hold.

Since the original S-OMP algorithm effectively gives \( K_0 \) distinct atoms from a dictionary \( A \) that best approximates the data matrix \( Y \) for \( K_0 \) iterations, we apply the general formulation even when the Hadamard operator is present. Recall that at every iteration \( k \), S-OMP measures the residual for each atom in \( A \) and creates an orthogonal projection with the highest correlation. If we adopt this scheme to color image setting, for every color channel \( c \), we can identify the index set \( \Lambda_k = [\Lambda_{r,k} \Lambda_{g,k} \Lambda_{b,k}] \) that give the highest correlation value:

\[
\lambda_k = \arg \max_{i=1,\ldots,M} \sum_{c \in \{r,g,b\}} w(c) \| R_c a_{c,i} \|_p
\]

where \( w(c) \) denotes the weight of each color channel and \( p \geq 1 \). After finding \( \lambda_k \), we modify the index setup to, \( \{\Lambda_{c,k}\} = \{\Lambda_{c,k-1} \cup \lambda_{c,k}\} \) for \( c \in \{r, g, b\} \). Thus, by finding the index set for the three distinct color channels based on some weight vector, we can create an orthogonal projection with each of the atoms in their corresponding color channels. The algorithm is summarized below in Algorithm 2.

3.4 Solving the Auxiliary Optimization Problem

The optimization problem in Equation (12) is convex and thus there exists a unique \( s^*_c \) for each \( c \in \{r, g, b\} \) that solves the problem. We use the \( \hat{S}' \) from the modified S-OMP to construct the initial membership matrix \( E \). Without any additional constraints, we obtain a closed form solution for \( s^*_c \).

**Problem.** The unique solution \( s^*_c \) for the optimization problem in Equation (12) is given as

\[
s^*_c = -H_c^{-1}(c_c - e_c \lambda)
\]

(13)

where \( H_c = A_c^T A_c, c_c = -A_c^T x_c, \) and \( \lambda = (e_c^T H_c^{-1} e_c)^{-1} e_c^T H_c^{-1} c_c \). Proof is shown in Appendix.
Algorithm 2: Color Channel S-OMP Algorithm

**Input:** $B \times 3ML$ dictionary matrix $A = [A_r, A_g, A_b]$ with $A_c = [A_{c,1}^T \cdots A_{c,M}^T] \in \mathbb{R}^{B \times ML}$ for $c \in \{r, g, b\}$, $B \times 3$ signal matrix $Y = [y_r, y_g, y_b]$, number of iterations $K$

**Initialization:** residual $R_0 = Y$, index set $\Lambda_0 = \emptyset$, iteration counter $k = 1$

while $k \leq K$ do

(1) Find the index of the atom that best approximates all residuals:
$$\lambda_{c,k} = \arg \max_{i=1, \ldots, M} \sum_{c \in \{r, g, b\}} w(c) \| R_{c,i}^T a_c \|_p, p \geq 1$$

(2) Update the index set $\Lambda_{c,k} = \Lambda_{c,k-1} \cup \{\lambda_{c,k}\}$

(3) Compute the orthogonal projector $P_{c,k} = (A_{c,k}^T A_{c,k})^{-1} A_{c,k}^T Y_c \in \mathbb{R}^{k}$ where $A_{c,k} \in \mathbb{R}^{B \times k}$ consists of the $k$ atoms in $A$ indexed in $\Lambda_{c,k}$ for each color channel $c \in \{r, g, b\}$

(4) Update the Residual Matrix $R_k = Y - [A_{c,k}^T P_{r,k} A_{c,k}^T P_{g,k} A_{c,k}^T P_{b,k}]$

(5) Increment $k$: $k \leftarrow k + 1$

end while

**Output:** Index set $\Lambda_c = \Lambda_{c,K}$, the sparse representation $\hat{S}'$ whose non-zero rows index for each color channel $c \in \{r, g, b\}$ by $\Lambda_c$ are the $K$ rows of the matrix $\left( A_{c,k}^T A_{c,k} \right)^{-1} A_{c,k}^T Y_c$

While Equation (15) gives the unique minimizer solution in Equation (12), we require a positivity constraint as negative coefficient in images would not have a physical meaning. Thus, we introduce constraint $s_c \geq 0$ to solve Equation (12). Note that with the additional inequality constraint, we forgo the closed form solution found above. Therefore, we use iterative methods to solve the problem. More specifically, we utilize interior point convex method for MATLAB’s `quadprog` in the Optimization Toolbox [23].

Therefore, with the addition of the positivity constraint, solving the auxiliary optimization problem has two implications. If we use $\hat{S}'$ obtained from any numerical solvers using any algorithms, the auxiliary optimization problem may improve the solution while guaranteeing sparsity. On the other hand, if we use S-OMP, the auxiliary problem only imposes positivity conditions since S-OMP results in the best selection of atoms in a given sparsity level.

### 3.5 Classification

After solving the auxiliary optimization problem to get an improved solution $\hat{s}_c$ for all $c \in \{r, g, b\}$ and obtaining the sparse coefficient matrix $S$, we compute the class-specific residual errors and finally identify the the class of the test image $X$ which gives the minimum residual as Equations (6) and (7):

$$\text{Class} (X) = \arg \min_{i=1, \ldots, M} \| X - A \delta_i (S) \|_F$$

where $\delta_i (S)$ is the matrix whose only non-zero entries are the same as those in $S$ associated with class $i$ in all color channels.
4 Simulation Results

The images that will be used to test the performance of our sparsity model are histopathological images provided by the Animal Diagnostics Laboratory (ADL) at Pennsylvania State University. Before we present the results, we start by describing the photomicrographs used for testing.

4.1 Histopathological Image Classification Process

The image database consists of photomicrographs of four different bovine organs: liver, lung, kidney, and spleen. The different conditions corresponding to each bovine organ are summarized in Table 1. The descriptions of the visual characteristics of each condition are shown below in Sections 4.1.1 through 4.1.4.

<table>
<thead>
<tr>
<th>Bovine Organ</th>
<th>Tissue Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lung</td>
<td>Healthy</td>
</tr>
<tr>
<td></td>
<td>Inflammation</td>
</tr>
<tr>
<td>Liver</td>
<td>Healthy</td>
</tr>
<tr>
<td></td>
<td>Inflammation</td>
</tr>
<tr>
<td></td>
<td>Necrosis</td>
</tr>
<tr>
<td>Kidney</td>
<td>Healthy</td>
</tr>
<tr>
<td></td>
<td>Inflammation</td>
</tr>
<tr>
<td>Spleen</td>
<td>Healthy</td>
</tr>
<tr>
<td></td>
<td>Inflammation</td>
</tr>
</tbody>
</table>

Table 1: Bovine Organ and Corresponding Conditions

4.1.1 Lung Photomicrographs

Some examples of lung images of the photomicrographs are shown in Figures 1 and 2. We note the fact that healthy lung images share a common trait of having white regions in the photomicrographs, whereas inflammation conditions do not.

Figure 1: Healthy Lung Tissue

Figure 2: Inflammatory Lung Tissue
4.1.2 Liver Photomicrographs

Next, the photomicrographs of liver are shown in Figures 3, 4 and 5. Although the liver tissue does not experience white regions like the lung, the images reflect a certain characteristic to each condition. The healthy liver tissue has nuclei, the blue dots that are uniformly scattered in the image. On the other hand, the inflammatory tissue has a large cluster of darker nuclei that are concentrated in the center. The main characteristic for necrotic tissue is the faded color of the nuclei and pale regions throughout the image.

![Figure 3: Healthy Liver Tissue](image1)

![Figure 4: Inflammatory Liver Tissue](image2)

![Figure 5: Necrotic Liver Tissue](image3)

4.1.3 Kidney Photomicrographs

Kidney photomicrographs exhibit similar conditions to the healthy and inflammatory conditions for liver tissue described in the previous subsection. The healthy kidney tissue exhibits uniform nuclei distributions throughout the image, whereas the inflammatory kidney tissues have a tendency for nuclei to be clustered in the center of the image.
4.1.4 Spleen Photomicrographs

Lastly, Spleen photomicrographs are shown above. We note that spleen tissues do not exhibit a significant difference in distribution of the nuclei in the image as the other organs. However, there are certain areas in the picture that reflect the tissue condition by some regions having paler portions of pink regions. By using the multispectral channels for classification, we expect to exploit this characteristic in our simulation.

4.2 Experimental Results

There are 50 images per class per organ with the exception of liver, for which only 25 images per class were provided. From the set of 50 images for each class of lung, kidney, and spleen, a subset of 45 training dictionary images were randomly selected. After combining the two dictionaries, we randomly chose a test image from the remaining images to test the performance of our algorithm. For liver, a subset of 20 images for each conditions, healthy, inflammatory, and necrotic, were randomly selected. Similarly, after combining the three dictionaries, we randomly chose a test image from the remaining set and performed classification.
To expedite classification process, the images were downsampled to $100 \times 75$. The process was repeated for 1000 trials for each class of the test image to remove any possible bias. The sparsity level was varied to observe the effects in the process. We compute the recognition rates for four approaches explained below in Table 2.

<table>
<thead>
<tr>
<th>Case</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grayscale</td>
<td>Grayscale S-OMP Output</td>
</tr>
<tr>
<td>Color - Uniform Weight</td>
<td>Color S-OMP with $w(c) = 1/3 \forall c$</td>
</tr>
<tr>
<td>Color - Weighted Sum</td>
<td>Color S-OMP with $w(c)$ based on color channel energy</td>
</tr>
<tr>
<td>Color - Auxiliary Output</td>
<td>Color - Weighted Sum with auxiliary problem</td>
</tr>
</tbody>
</table>

Table 2: Explanation of Simulation Methods

The classification results are shown in Figure 10 and 11. We first note that with $K = 2$, which corresponds to less than 5% of the number of atoms, we have poor results, as all of the recognition rates are near 0% or 100%. This can be interpreted as the initial bias resulting from our dictionary. Since S-OMP picks the index that best approximates all the residuals, this bias is mitigated by increasing the number of atoms in the classification process, which is evident as we increase the number of atoms $K$. Note that with $K = 20$, we are using 20% and 33% of dictionary elements for non-liver and liver tissues respectively, which still is considered sparse. For further analysis, we proceed with $K = 20$. The confusion matrices for the corresponding organs are shown in Tables 3 through 6.

Figure 10: Classification Results for Lung and Liver Tissue
Figure 11: Classification Results for Kidney and Spleen Tissue

### Table 3: Confusion Matrix: Lung

<table>
<thead>
<tr>
<th>Class</th>
<th>Healthy</th>
<th>Inflamm.</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthy</td>
<td>0.922</td>
<td>0.078</td>
<td>Grayscale</td>
</tr>
<tr>
<td>Inflamm.</td>
<td>0.233</td>
<td>0.767</td>
<td>Grayscale</td>
</tr>
<tr>
<td>Necrosis</td>
<td>0.163</td>
<td>0.837</td>
<td>Color-Aux</td>
</tr>
</tbody>
</table>

### Table 4: Confusion Matrix: Liver

<table>
<thead>
<tr>
<th>Class</th>
<th>Healthy</th>
<th>Inflamm.</th>
<th>Necrosis</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthy</td>
<td>0.739</td>
<td>0.261</td>
<td></td>
<td>Grayscale</td>
</tr>
<tr>
<td>Inflamm.</td>
<td>0.385</td>
<td>0.615</td>
<td></td>
<td>Color-Aux</td>
</tr>
<tr>
<td>Necrosis</td>
<td>0.306</td>
<td>0.694</td>
<td></td>
<td>Color-Aux</td>
</tr>
</tbody>
</table>

### Table 5: Confusion Matrix: Kidney

<table>
<thead>
<tr>
<th>Class</th>
<th>Healthy</th>
<th>Inflamm.</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthy</td>
<td>0.706</td>
<td>0.294</td>
<td>Grayscale</td>
</tr>
<tr>
<td>Inflamm.</td>
<td>0.634</td>
<td>0.366</td>
<td>Grayscale</td>
</tr>
<tr>
<td>Necrosis</td>
<td>0.955</td>
<td>0.045</td>
<td>Color-Aux</td>
</tr>
</tbody>
</table>

### Table 6: Confusion Matrix: Spleen

<table>
<thead>
<tr>
<th>Class</th>
<th>Healthy</th>
<th>Inflamm.</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthy</td>
<td>0.739</td>
<td>0.261</td>
<td>Grayscale</td>
</tr>
<tr>
<td>Inflamm.</td>
<td>0.385</td>
<td>0.615</td>
<td>Color-Aux</td>
</tr>
<tr>
<td>Necrosis</td>
<td>0.306</td>
<td>0.694</td>
<td>Color-Aux</td>
</tr>
</tbody>
</table>

Now, let us consider only the outputs from color output from the auxiliary problem and grayscale classification. The recognition rate for these two methodology is shown below in Figure 12 along with each classification’s 95% confidence interval. Note that we use Clopper-Pearson method to generate the confidence interval [24]. The classification method using color channel has higher recognition rate than using grayscale images. The difference is mostly reflected on classification of spleen, where the classes did not exhibit differences in nuclei distribution. Clearly, the performance difference between the two methods show the advantage of exploiting color information. With the exception of Lung, classification using color channels provide a better performance than guessing (1/2 for binary classification, 1/3 for ternary classification).
Lastly, we compare the recognition rates from our color output with the rate obtained by a classification method of using feature extraction and support vector machine (SVM) [25]. The recognition rate with the corresponding 95% confidence interval is shown below in Figure 13. While feature extraction method dominates our method most of the time, we observe that our method does better for kidney tissues. This may result from the fact that kidney tissues have a more uniform structure than other organ tissues. Since our methodology is a form of example-based learning, we would expect better results with better dictionary elements.

![Grayscale-RGB 95% Confidence Interval](image1)

**Figure 12: 95% Confidence Interval for Grayscale and Color-Aux**

![Feature Extraction+SVM and S-OMP 95% Confidence Interval](image2)

**Figure 13: 95% Confidence Interval for Feature Extraction and S-OMP**
5 Conclusion

The classification approach introduced in this paper utilizes multispectral channels for a sparsity model to classify images. To adopt the original grayscale sparsity model proposed in [9] to color images, we extend joint sparsity models suggested in [10]. Then, since the original problem was non-convex, we use a matching pursuit algorithm (S-OMP) to approximate the problem. The initially obtained $S$ from S-OMP was improved by solving an additional optimization problem that enforces positivity while guaranteeing sparsity. The reconstruction parameters were then used to calculate the residual norm for each of the different classes. Finally, the class with minimum residual was chosen.

For testing purposes, we use bovine organ database provided by the Pennsylvania State University Animal Diagnostic Laboratory. For every organ and class, subsets of the images were chosen to be training samples that would be the dictionary of our model. Then, a test image was chosen and our algorithm performed classification. Our model was tested for $N = 1000$ trials. The recognition rate was higher for using color channels than grayscale by exploiting additional information. The difference was mostly noticeable where the deterministic information for classification was present in the color channels, not the overall structure of the image. Lastly, our performance was compared with classification with feature extraction and support vector machine obtained from [25]. While our algorithm showed better performance for the recognition rate for kidney tissue, feature extraction/SVM method dominated for the other organ tissues. This suggests the importance of processing the images by using feature extraction or contrast agents in medical imaging. While for histopathological images our algorithm does not show a robust performance, using multispectral channel does reflect a stronger performance than single channel.

5.1 Suggestions for Further Work

Our model formulation was an approach to adapt the grayscale sparsity model to multispectral channels. Although different methods using PCA/ICA have been already developed, we concentrate on sparse model framework. Since the Hadamard operator from $S$ to $S'$ in Equation (9) impose a complexity in using previous methods for solving sparse problems, further work to relax the operator is suggested. Also, instead of using S-OMP, other approximation methods could be used to show the performance in multichannel operations.

For the histopathological image classification, possibilities for future improvements are suggested. Since the performance of S-OMP is depended upon the strength of the dictionary, diverse images would be needed to ensure sufficiently high recognition rate. In the case of medical imaging, where low Type I error is essential, this may require thousands of images. Also, the analysis of performance gain of using preprocessed medical images by feature extraction or contrast agents in dictionary with sparse models is suggested.
5.2 Acknowledgments

I would first like to thank Professor Vishal Monga and Umamahesh Srinivas for the helpful discussions and assistance with the project. Also, I would like to thank the EEREU program at Pennsylvania State University for giving me the opportunity to participate in research. Furthermore, I would like to give great thanks to Professor Saleem Kassam and Santosh Venkatesh at Penn for assisting my studies. Lastly, I would like to acknowledge Mu Li and Bosung Kang for making my experience at Information Processing and Algorithms Lab (iPAL) and Pennsylvania State University memorable. This material is based upon work supported by the National Science Foundation under Grant No. EEC-1062984.

6 Appendix

Problem. The unique solution $s^*_c$ for the optimization problem in Equation (12) is given as

$$s^*_c = -H_c^{-1} (c_c - e_c \lambda)$$

where $H_c = A_c^T A_c$, $c_c = -A_c^T x_c$, and $\lambda = (e_c^T H_c^{-1} e_c)^{-1} e_c^T H_c^{-1} c_c$.

Proof. The proof is based from an example in [26]. We start by examining the quadratic programming problem

$$\begin{align*}
\text{minimize} & \quad \frac{1}{2} \|x\|^2 + c^T x \\
\text{subject to} & \quad Ax = 0
\end{align*}$$

where $c \in \mathbb{R}^n$ and $A$ is an $m \times n$ matrix of rank $m$. Since adding a constant term $\frac{1}{2} \|c\|^2$ does not affect our cost function, we rewrite the problem as

$$\begin{align*}
\text{minimize} & \quad \frac{1}{2} \|c + x\|^2 \\
\text{subject to} & \quad Ax = 0
\end{align*}$$

which is the projection of the vector $-c$ onto the subspace $X = \{x | Ax = 0\}$. It can be seen that the unique solution to the problem above is given as the orthogonal projector of the vector:

$$x^* = - \left( I - A^T (AA^T)^{-1} A \right) c$$

(15)
With the orthogonal projector ready consider a more general quadratic program

$$\begin{align*}
\text{minimize} & \quad \frac{1}{2}x^T H x + c^T x \\
\text{subject to} & \quad Ax = 0
\end{align*}$$

where $H$ is a positive definite symmetric matrix. By introducing the vector $y = H^{1/2}x$, we transform this problem to

$$\begin{align*}
\text{minimize} & \quad \frac{1}{2} \|y\|^2 + \left( H^{-1/2} c \right)^T y \\
\text{subject to} & \quad AH^{-1/2} y = 0
\end{align*}$$

Using Equation (15) the solution of this problem is

$$y^* = - \left( I - H^{-1/2} A^T (AH^{-1} A^T)^{-1} AH^{-1/2} \right) H^{-1/2} c$$

thus by passing through the transformation $x^* = H^{-1/2} y^*$, the optimal solution of problem of Equation (16) is given as

$$x^* = - H^{-1} \left( c - A^T \lambda \right)$$

where $\lambda$ is given by

$$\lambda = (AH^{-1} A^T)^{-1} AH^{-1} c$$
References


