Simultaneous Sparsity Model for Histopathological Image Representation and Classification

Umamahesh Srinivas, Student Member, IEEE, Hojjat Seyed Mousavi, Student Member, IEEE, Vishal Monga, Senior Member, IEEE, Arthur Hattel, and Bhushan Jayarao

Abstract—The multi-channel nature of digital histopathological images presents an opportunity to exploit the correlated color channel information for better image modeling. Inspired by recent work in sparsity for single channel image classification, we propose a new simultaneous Sparsity model for multi-channel Histopathological Image Representation and Classification (SHIRC). Essentially, we represent a histopathological image as a sparse linear combination of training examples under suitable channel-wise constraints. Classification is performed by solving a newly formulated simultaneous sparsity-based optimization problem. A practical challenge is the correspondence of image objects (cellular and nuclear structures) at different spatial locations in the image. We propose a robust locally adaptive variant of SHIRC (LA-SHIRC) to tackle this issue. Experiments on two challenging real-world image data sets: (i) mammalian tissue images acquired by pathologists of the Animal Diagnostics Lab (ADL) at Pennsylvania State University, and (ii) human intraductal breast lesions, reveal the merits of our proposal over state-of-the-art alternatives. Further, we demonstrate that LA-SHIRC exhibits a more graceful decay in classification accuracy against the number of training images which is highly desirable in practice where generous training per class is often not available.

Index Terms—Classification, histopathological image analysis, multichannel images, sparse representation.

I. INTRODUCTION

The advent of digital pathology [1] has ushered in an era of computer-assisted diagnosis and treatment of medical conditions based on the analysis of medical images. Of active research interest is the development of quantitative image analysis tools to complement the efforts of radiologists and pathologists towards better disease diagnosis and prognosis. This research thrust has been fueled by a variety of factors, including the availability of large volumes of patient-related medical data, dramatic improvements in computational resources (both hardware and software), and algorithmic advances in image processing, computer vision and machine learning theory. An important emerging sub-class of problems in medical imaging pertains to histopathological images. Recent work has identified the nascentness of quantitative research in histopathological image analysis and classification, and its potential for important practical applications [2]–[7]. Whole slide digital scanners process tissue slides to generate these digital images. Examples of histopathological images are shown in Figs. 6 and 9. It is evident that these images carry rich structural information, making them invaluable for the diagnosis of many diseases including cancer [6], [8]–[10].

Pathologists often look for visual cues at the nuclear and cellular level in order to categorize a tissue image as either healthy or diseased. Motivated by this, a variety of low-level image features have been developed based on texture, morphometric characteristics (shape and spatial arrangement) and image statistics. The gray level co-occurrence matrix by Haralick et al. [11] estimates the texture of an image in terms of the distribution of co-occurring pixel intensities at specified offset positions. Morphological image features have been used in medical image segmentation for detection of vessel-like patterns [13]. Image histograms are a popular choice of features for medical imaging [14]. Wavelet features have been deployed for prostate cancer diagnosis in [15]. Esgiar et al. [16] have captured the self-similarity in colon tissue images using fractal-based features. Tabesh et al. [17] have combined color, texture and morphometric features for prostate cancer diagnosis. Doyle et al. [18] introduced graph-based features using Delaunay triangulation and minimum spanning trees to exploit spatial structure. In [6], a hybrid classification model that combines structural and statistical information has been developed. Orlov et al. [19], [20] have recently proposed a multi-purpose feature set that aggregates transform domain coefficients, image statistics and texture information. Experimental success in many different classification problems has demonstrated the versatility of this feature set. It must be mentioned that all the features discussed above are applicable broadly for image analysis and have been particularly successful in medical imaging. For classification, these features are combined with powerful classifiers such as support vector machines (SVMs) [14], [21] and boosting [22], [23]. A comprehensive discussion of features and classifiers for histopathological analysis is provided in [4].

A. Motivation and Challenges

While histopathology shares some commonalities with other popular imaging modalities such as cytology and radiology, it also exhibits two principally different characteristics [24] that pose challenges to image analysis. First, histopathological images are invariably multi-channel in nature (commonly using three color channels - red, green and blue (RGB)). Key
geometric information is spread across the color channels. It is well known that color information in the hematoxylin-eosin (H&E) stained slides is essential to identify the discriminative image signatures of healthy and diseased tissue [7], [25], [26]. Specifically, the nuclei assume a bluish tinge due to hematoxylin, while the cytoplasmic structures and connective tissue appear red due to eosin. As seen from Fig. 6, there is a higher density of nuclei in diseased tissue. Typically in histopathological image analysis, features are extracted from each color channel of the images [17], and the classifier decisions based on the individual feature sets are then fused for classification. Alternately, only the luminance channel information - image edges resulting mainly from illumination variations - is considered [26]. The former approach ignores the inherent correlations among the RGB channels while the latter strategy fails to exploit chrominance channel geometry, i.e. the edges and image textures caused by objects with different chrominance.

The second challenge posed by histopathology is the relative difficulty in obtaining good features for classification due to the geometric richness of tissue images. Tissues from different organs have structural diversity and often, the objects of interest occur at different scales and sizes [4]. As a result, features are usually customized for specific classification problems, most commonly cancer of the breast and prostate.

In this paper, we address both these challenges through a novel simultaneous sparsity model inspired by recent work using sparse representations for image classification (SRC) [27]. Given a sufficiently diverse collection of training images from each class, Wright et al. [27] conjecture that any other image from the same class can be expressed approximately as a linear combination of those training images. So, each test image has a sparse representation in terms of a basis matrix or dictionary comprising training images from all classes. This linear additive model alleviates the challenge of designing sophisticated task-specific features. Additionally, the class-specific design of such dictionaries enables class assignment via a simple comparison of class-specific reconstruction errors. The robustness of sparse features to real-world imaging distortions has led to their widespread use in application domains such as remote sensing (hyperspectral imaging [28], synthetic aperture radar [29]) and biometrics (face recognition [27], [30]). SRC has also been proposed for single-channel medical images, in cervigram segmentation [31], [32] and colorectal polyp and lung nodule detection [33]. To the best of our knowledge, ours is the first discriminative sparsity model for multi-channel histopathological images.

B. Our Contributions

The relevance of color information for image classification tasks has been identified previously [35], [36]. We propose a new sparsity model that recognizes and exploits the diversity of information in multiple color channels of histopathological images. Specifically, our key contributions are listed next.

1) **Simultaneous sparsity model for classification.** Essentially, our simultaneous Sparsity model for multi-channel Histopathological Image Representation and Classification (SHIRC) extends the standard SRC approach [27]–[33] by designing three color dictionaries, corresponding to the RGB channels. Each multi-channel histopathological image is represented as a sparse linear combination of training examples under suitable channel-wise constraints, which capture color correlation information. The constraints agree with intuition since a sparse linear model for a color image necessitates identical models for each of its constituent color channels with no cross-channel contributions.

2) **Novel optimization problem.** Our approach considers a multi-task scenario that is qualitatively similar to the visual classification problem addressed very recently by Yuan et al. [37]. In [37], three types of image features - containing texture, shape and color information - are extracted from images and a joint sparsity model is proposed to classify the images. The joint (simultaneous) sparsity model employed in [37] and the one we develop however have some differences. First, [37] does not consider the problem of multi-channel or color image modeling. Second and more crucially, the cost function in [37] is a sum of reconstruction error terms from each of the feature dictionaries which results in the commonly seen row sparsity structure on the sparse coefficient matrix. The resulting optimization problem is solved using the popular Accelerated Proximal Gradient (APG) method [38]. In our work however, to conform to imaging physics, we introduce color channel-specific constraints on the structure of the sparse coefficients, which do not directly conform to row sparsity, leading to a new optimization problem. This in turn requires a modified greedy matching pursuit approach to solve the problem (see Appendix A).

3) **Locally adaptive SHIRC: A practical solution to handle image correspondence.** As discussed earlier, feature design in medical imaging is guided by the object-based perception of pathologists. Depending on the type of tissue, the objects could be nuclei, cells, glands, lymphocytes, etc. Crucially it is the presence or absence of these local objects in an image that matters to a pathologist; their absolute spatial location matters much less. As a result, the object of interest may be present in the test image as well as the representative training images, albeit at different spatial locations, causing a seeming breakdown of the image-level SHIRC. This scenario can occur in practice if the acquisition process is not carefully calibrated.

So we infuse the SHIRC with a robust locally adaptive flavor by developing a Locally Adaptive SHIRC (LA-SHIRC). Crucially, we rely on the *pathologist’s insight* to carefully select multiple local regions that contain these objects of interest from training as well as test images and use them in our linear sparsity model instead of the entire images. Local image features often possess better discriminative power than image-level features.
II. BACKGROUND: SPARSE REPRESENTATION-BASED IMAGE CLASSIFICATION (SRC)

Suppose that there are \( K \) different image classes (corresponding to categories of medical conditions), labeled \( C_1, \ldots, C_K \). Let there be \( N_i \) training samples (each in \( \mathbb{R}^n \)) corresponding to class \( C_i, i = 1, \ldots, K \). It is understood that each sample is the vectorized version of the corresponding grayscale (or single channel) image. The training samples corresponding to class \( C_i \) can be collected in a matrix \( D_i \in \mathbb{R}^{n \times N_i} \), and the collection of all training samples is expressed using the matrix:

\[
D = [D_1 \ D_2 \ \ldots \ D_K],
\]

where \( D \in \mathbb{R}^{n \times T} \), with \( T = \sum_{k=1}^{K} N_k \). A new test sample \( y \in \mathbb{R}^n \) can be expressed as a sparse linear combination of the training samples:

\[
y \simeq D \alpha_1 + \ldots + D_K \alpha_K = D \alpha,
\]

where \( \alpha \) is ideally expected to be a sparse vector (i.e., only a few entries in \( \alpha \) are nonzero). The classifier seeks the sparsest representation by solving the following problem:

\[
\hat{\alpha} = \underset{\alpha}{\arg \min} \|\alpha\|_0 \quad \text{subject to} \quad \|D \alpha - y\|_2 \leq \varepsilon, \tag{3}
\]

where \( \|\cdot\|_0 \) denotes the number of nonzero entries in the vector and \( \varepsilon \) is a suitably chosen reconstruction error tolerance. The problem in (3) can be solved by greedy pursuit algorithms [40], [41]. Once the sparse vector is recovered, the identity of \( y \) is given by the minimal class-specific reconstruction residual:

\[
\text{Class}(y) = \underset{i}{\arg \min} \|y - D \delta_i(\hat{\alpha})\|_2, \tag{4}
\]

where \( \delta_i(\alpha) \) is a vector whose only nonzero entries are the same as those in \( \alpha \) which are associated with class \( C_i \).

Rooted in optimization theory, the robustness of the sparse feature to real-world image distortions like noise and occlusion has led to its widespread application in practical classification tasks. Modifications to (3) include relaxing the non-convex \( l_0 \)-term to the \( l_1 \)-norm [42] and introducing regularization terms to capture physically meaningful constraints [32].

In many scenarios, we have access to multiple sets of measurements that capture information about the same image. The SRC model is extended to incorporate this additional information by enforcing a common support set of training images for the \( T \) correlated test images \( y_1, \ldots, y_T \):

\[
Y = [y_1 \ y_2 \ \ldots \ y_T] = [D \alpha_1 \ D \alpha_2 \ \ldots \ D \alpha_T] = D \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_T \end{bmatrix} = DS.
\]

(5)

The vectors \( \alpha_i, i = 1, \ldots, T \), all have non-zero entries at the same locations, albeit with different weights, leading to the recovery of a sparse matrix \( S \) with only a few nonzero rows:

\[
\hat{S} = \underset{S}{\arg \min} \|Y - DS\|_F \quad \text{subject to} \quad \|S\|_{\text{row,0}} \leq K_0, \tag{6}
\]

where \( \|S\|_{\text{row,0}} \) denotes the number of non-zero rows of \( S \) and \( \|\cdot\|_F \) is the Frobenius norm. The greedy Simultaneous Orthogonal Matching Pursuit (SOMP) [43] algorithm and convex relaxations of the row-sparsity norm [44] have been proposed to solve the non-convex problem in (6). Example real-world manifestations of this multi-variante scenario occur...
in the form of spatially local pixels for hyperspectral target classification [28], multiple feature sets for automatic image annotation [45], kernel features for object categorization, or as query images for video-based face recognition [37].

III. CONTRIBUTIONS

A. SHIRC: A simultaneous Sparsity model for Histopathological Image Representation and Classification

Section II has identified the central analytical formulation underlying the simultaneous sparsity methods in literature. Typically a single dictionary $D$ is used, as in [28]. In other cases, each event is in fact characterized by multiple heterogeneous sources, resulting in multi-task versions of SRC [37], [45]. Although different dictionaries are used for the different sources, the issue of correlation among different representations of the same image is not thoroughly investigated. Our contribution in this paper is an example of multi-task classification, with separate dictionaries designed from the RGB channels of histopathological images.

For ease of exposition, we consider the binary classification problem of classifying images as either healthy or diseased. $D_h$ and $D_d$ indicate the training dictionaries of healthy and diseased images respectively. A total of $N$ training images are chosen. We represent a color image as a matrix $Y := [y^r \ y^g \ y^b] \in \mathbb{R}^{n \times 3}$, where the superscripts $r,g,b$ correspond to the RGB color channels respectively. The dictionary $D$ is redefined as the concatenation of three color-specific dictionaries, $D := [D^r \ D^g \ D^b] \in \mathbb{R}^{n \times 3N}$. Each color dictionary $D^c, c \in \{r,g,b\}$, is the concatenation of sub-dictionaries from both classes belonging to the $c$-th color channel:

$$D^c := [D^c_h \ D^c_d], c \in \{r,g,b\}. \quad (7)$$

$$D := [D^r \ D^g \ D^b] = [D^r_h \ D^r_d \ D^g_h \ D^g_d \ D^b_h \ D^b_d]. \quad (8)$$

The color dictionaries are designed to obey column correspondence, i.e., the $i$-th column from each of the color dictionaries $D^c$ taken together correspond to the $i$-th training image. Fig. 1 shows the arrangement of training images into channel-specific dictionaries. A test image $Y$ can now be represented as a linear combination of training samples as follows:

$$Y = DS = \begin{bmatrix} D^r_h & D^r_d & D^g_h & D^g_d & D^b_h & D^b_d \end{bmatrix} \begin{bmatrix} \alpha^r \ \alpha^g \ \alpha^b \end{bmatrix}, \quad (9)$$

where the coefficient vectors $\alpha^c \in \mathbb{R}^{3N}, c \in \{r,g,b\}$, and $S = [\alpha^r \ \alpha^g \ \alpha^b] \in \mathbb{R}^{3N \times 3}$.

A natural question to ask at this juncture is: why do we need a model that permits a separate dictionary per channel? A naive alternative would be to use a single color dictionary, wherein the channels of each color image are stacked into a single column. Let $x = [x^r \ x^g \ x^b]^T \in \mathbb{R}^{3n}$ denote one such training image. Define the matrix $W$ as follows:

$$W := [w_r I \ w_g I \ w_b I] \in \mathbb{R}^{n \times 3n},$$

where $I \in \mathbb{R}^{n \times n}$ is the identity matrix. $\hat{x} = Wx = w_r x_r + w_g x_g + w_b x_b \in \mathbb{R}^{n}$ is the grayscale version of $x$ when the weights $w_r, w_g, w_b$ are chosen appropriately. Now consider a training dictionary $A$ obtained by stacking the $N$ vectorized color
images $\mathbf{x}$ into a $3n \times N$ matrix:

$$A := \begin{bmatrix} D^r \\ D^g \\ D^b \end{bmatrix}$$

A test image $\mathbf{y}$ can now be written as:

$$\mathbf{y} = A\mathbf{a},$$

where $\mathbf{a} \in \mathbb{R}^N$ is the coefficient vector. Applying the transformation matrix $W$ to both sides of the above equation,

$$Wy = WA\mathbf{a} \Rightarrow \hat{y} = \hat{A}\mathbf{a},$$

which is the linear representation model for the grayscale (single channel) case. Crucially, the coefficient vector $\mathbf{a}$ has not changed. This shows that the model which uses a single vector per color image is identical to the grayscale image model in SRC.

It remains to be justified that a separate coefficient per channel is physically meaningful. There is high variability in the staining process inherent to histopathological imaging [4]. This can lead to training and test images with different gains in the color channels. Using a coefficient per channel incorporates color balancing in the model.

A closer investigation of $\mathbf{S}$ reveals interesting characteristics:

1) It is reasonable to assume that the $c$-th channel of the test image (i.e. $\mathbf{y}'$) can be represented by the linear span of the training samples belonging to the $c$-th channel alone (i.e. only those training samples in $D^c$). So the columns of $\mathbf{S}$ ideally have the following structure:

$$\mathbf{a}' = \begin{bmatrix} \mathbf{a}'_h \\ \mathbf{a}'_g \\ \mathbf{a}'_b \end{bmatrix}, \quad \mathbf{a}'_c = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{a}'_h = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix},$$

where $\mathbf{0}$ denotes the conformal zero vector. In other words, $\mathbf{S}$ exhibits block-diagonal structure.

2) Each color channel representation $\mathbf{y}'$ of the test image is in fact a sparse linear combination of the training samples in $D^c$. Suppose the image belongs to class $h$ (healthy); then only those coefficients in $\mathbf{a}'_c$ that correspond to $D^c_h$ are expected to be non-zero.

3) The locations of non-zero weights of color training samples in the linear combination exhibit one-to-one correspondence across channels. If the $j$-th training sample in $D^c$ has a non-zero contribution to $\mathbf{y}'$, then for $c \in \{g, b\}$, $\mathbf{y}'$ has non-zero contribution from the $j$-th training sample in $D^c$.

**B. Optimization Problem**

This immediately suggests a joint sparsity model similar to (6). However, the row sparsity constraint leading to the SOMP solution is not obvious from this formulation. Instead, we introduce a new matrix $\mathbf{S}' \in \mathbb{R}^{N \times 3}$ as the transformation of $\mathbf{S}$ with the redundant zero coefficients removed:

$$\mathbf{S}' = \begin{bmatrix} \mathbf{α}'_h \\ \mathbf{α}'_g \\ \mathbf{α}'_b \end{bmatrix},$$

This is possible by first defining $H \in \mathbb{R}^{3N \times 3}$ and $J \in \mathbb{R}^{N \times 3N}$,

$$H = \begin{bmatrix} 1_N & 0 & 0 \\ 0 & 1_N & 0 \\ 0 & 0 & 1_N \end{bmatrix}, \quad J = [I_N \quad I_N \quad I_N],$$

where $1_N \in \mathbb{R}^N$ is the vector of all ones, and $I_N$ denotes the $N$-dimensional identity matrix. Now,

$$\mathbf{S}' = J(H \circ \mathbf{S}),$$

where $\circ$ denotes the Hadamard product, $(H \circ \mathbf{S})_{ij} = h_{ij} s_{ij}$ $\forall i, j$. Finally, we formulate a sparsity-enforcing optimization problem that is novel to the best of our knowledge:

$$\hat{\mathbf{S}}' = \arg\min \|\mathbf{S}'\|_{row,0} \text{ subject to } \|\mathbf{Y} - DS\|_F \leq \epsilon.$$  \hspace{1cm} (13)

Solving the problem in (13) presents a challenge in that a straightforward application of SOMP [43] is not possible due to the non-invertibility of the Hadamard operator. We have developed a greedy algorithmic modification of SOMP that fares well in practice. In the interest of presenting all our contributions in a logically motivated sequence, the analytical details of the algorithm are presented in Appendix A.

The final classification decision is made by comparing the class-specific reconstruction errors to a threshold $\tau$:

$$R(\mathbf{Y}) = \left\| \mathbf{Y} - D_h \hat{\mathbf{S}}_h \right\|_2 - \left\| \mathbf{Y} - D_d \hat{\mathbf{S}}_d \right\|_2 \geq \tau,$$  \hspace{1cm} (14)

where $\hat{\mathbf{S}}_h$ and $\hat{\mathbf{S}}_d$ are matrices whose entries are those in $\hat{\mathbf{S}}$ associated with $D_h$ and $D_d$ respectively.

Shown in Fig. 5 is an illustration of the reconstructed versions of the original image, which is in fact that of a cancerous (DCIS) breast lesion from the IBL data set. The reconstructed images using training from the healthy and inflammatory classes are separately shown. It is clear that reconstruction from the true class (DCIS in this case) more closely resembles the original image than the reconstruction from the UDH class. The reconstruction residuals for the two classes are 0.31 and 0.81 respectively, showing that the two classes are clearly separable using this metric.

**Multi-class problems:** In the ensuing experimental validation, we consider binary (healthy versus inflammatory/diseased) classification problems from two different data sets. In general,
Fig. 4. LA-SHIRC: Locally adaptive variant of SHIRC. The yellow boxes indicate local objects of interest such as cells and nuclei. The new dictionary $\mathbf{D}$ is built using multiple local blocks from each training image. In every test image, the local objects are classified using the simultaneous sparsity model and their decisions are fused for overall image-level classification.

(a) Original DCIS test image, (b) Reconstructed image from UDH dictionary; reconstruction error = 0.81, (c) Reconstructed image from DCIS dictionary; reconstruction error = 0.31.

A more challenging problem is that of grading inflammations. Our approach extends to such a $K$-class scenario in a straightforward manner by incorporating additional class-specific dictionaries in $\mathbf{D}$. The classification rule is then modified as follows:

$$\text{Class}(\mathbf{Y}) = \arg \min_{k=1, \ldots, K} \| \mathbf{Y} - \mathbf{D} \delta_k(\hat{\mathbf{S}}) \|_F,$$

(15)

where $\delta_k(\hat{\mathbf{S}})$ is the matrix whose only non-zero entries are the same as those in $\hat{\mathbf{S}}$ associated with class $C_k$.

C. LA-SHIRC: Locally Adaptive SHIRC

Some histopathological image collections present a unique practical challenge in that the presence or absence of desired objects (e.g., cells, nuclei) in images is more crucial - compared to their actual locations - for pathologists for disease diagnosis. Consequently, if these discriminative objects (sub-images) are not in spatial correspondence in the test and training images, it would seem that SHIRC cannot handle this scenario. Fig. 2 illustrates this for sample images from the IBL data set.

This issue can be handled practically to some extent by careful pre-processing that manually segments out the objects of interest for further processing [10]. However this approach causes a loss in contextual information that is also crucial for pathologists to make class decisions. We propose a robust algorithmic modification of SHIRC, known as Locally Adaptive SHIRC (LA-SHIRC), to address this concern.

It is well known that local image features are often more useful than global features from a discriminative standpoint [39]. This also conforms with the intuition behind pathologists’ decisions. Accordingly, we modify SHIRC to classify local image objects instead of the global image. Several local objects of the same dimension (vectorized versions lie in $\mathbb{R}^m$, $m \ll n$) are identified in each training image based on the recommendation of the pathologist. Fig. 3 shows individual cells from the IBL data set. In fact, these obvious differences in cell structure have been exploited for classification [10] by designing morphological features such as cell perimeter and ratio of major-to-minor axis of the best-fitting ellipse.

The dictionary $\mathbf{D}$ in SHIRC is now replaced by a new dictionary $\hat{\mathbf{D}}$ that comprises the local blocks. In Fig. 4, the yellow boxes indicate the local regions. Assuming $N$ full-size training images, the selection of $B$ local blocks per image...
results in a training dictionary of size $NB$. Note that even for fixed $N$, the dictionary $D \in \mathbb{R}^{m \times NB}$ has more samples (or equivalently, leads to a richer representation) than $D \in \mathbb{R}^{n \times N}$. Therefore, a test image block is expressed as a sparse linear combination of all local image blocks from the training images. $B$ blocks are identified in every test image, and a class decision is obtained for each such block by solving (13) but with the dictionary $D$. Finally the identity of the overall image is decided by combining the local decisions from its constituent blocks.

D. Local Block Selection: A Practical Approach

Fig. 7 shows the implementation pipeline of LA-SHIRC for any general histopathological image data set. The figure is divided into two phases: training and test. The input to the training phase is a collection of labeled training images. We also benefit from the pathologist’s insight to identify the discriminative local regions in each image. For example, in Fig. 3, we see the differences between the (local) cell structures from the two classes. Our first goal is to segment out multiple such regions from all training images. In the process of identifying the local regions, the pathologist estimates the relative size of such regions in comparison with the size of the whole image. To illustrate, in the IBL data set (Fig. 2), a possible estimate could be that the local regions roughly occupy 5% of the entire image (or say, a block of $40 \times 40$ pixels). With this information, we can choose our local blocks using either of two strategies.

We can incorporate an informed image segmentation pre-processing step to identify the local regions. In our experiments on the IBL data set, we identify that the local regions can be captured very well using a classical morphology-based blob detection technique [12]. Using the estimate of the local region’s size, we choose an appropriate blob dimension and identify all such blobs (or cells) in each of the training images. Since each image in our data set contains many cells, we extract a reasonable number per training image, typically about 10. More generally, we can leverage state-of-the-art segmentation algorithms that have been designed specifically for certain image data sets.

Alternately, we fix the size of a rectangular block based on the cell size estimate and adopt a blind random tiling strategy to select multiple overlapping blocks in each training image. Fig. 8 an illustration. This idea is in fact inspired by similar strategies employed for robust face recognition [46] and secure image hashing [47]. The goal is to uniformly select from different regions of the image; choosing overlapping regions ensures that most of the image is covered. Here, the number of such blocks per training image is a function of how densely the discriminative regions are distributed in the overall image, a decision that is again made by the pathologist. It is reasonable to expect that classification performance will improve as the number of blocks in the test image increases, albeit possibly at the cost of more computation. While both
strategies are viable, preference for either is based on factors such as: (i) availability of custom segmentation tools for a particular data set, and (ii) the additional computation incurred in the selection (and classification) of a larger number of blocks to make the performance of the random tiling strategy comparable to the informed approach.

In both strategies, the dictionary $\mathcal{D}$ is constructed from all these local regions from all training images. It must be emphasized that only the structural information about the local blocks is captured in the dictionary; where they come from (spatially) in the training images, and which training images they come from, are not relevant.

The block selection process for a new test image is consistent with the strategy chosen for the training images. It is reasonable to believe that the relative scale of the local blocks is similar to the scale in training images - this can be calibrated fairly accurately during the image acquisition process; if not, the test images can be suitably resized. As a result, we now have an instantiation of our SHIRC framework but at the local block level.

**Benefits of LA-SHIRC:**

- We represent a single local block from a test image as a sparse linear combination of all local blocks from all training images. Therefore, this approach obviates the need for image registration at the global (whole image) level, which is in itself a challenging problem.
- We classify several individual blocks from the test image separately and combine the decisions as discussed next. The number of local blocks in the test image should be comparable to the number used in the training process.
- The *adaptive* term in the name LA-SHIRC indicates the flexibility offered by the algorithm in terms of choosing the number of blocks $B$ and their dimension $m$. Objects of interest in histopathological image analysis exist at different scales [4] and the tunability of LA-SHIRC makes it amenable to a variety of histopathological classification problems. LA-SHIRC satisfactorily handles the issue of spatial correspondence of objects. Additionally, as will be demonstrated through experiments in Section IV-D, it ensures high classification accuracy even with a small number of (global) training images. This has high practical relevance since generous number of training histopathological images per condition (healthy/diseased) may not always be available.

**Decision fusion:** A common way of combining local class decisions is majority voting. Suppose $Y_i, i = 1, \ldots, B$, represent the $B$ local blocks from image $Y$. Then,

$$\text{Class}(Y) = \max_{k=1,\ldots,K} |\{i : \text{Class}(Y_i) = k\}|, i = 1, \ldots, B,$$

where $|\cdot|$ denotes set cardinality and Class$(Y_i)$ is determined by (15). This scheme suffers from the drawback that it is not a soft measure, i.e. there is no indicator of the degree of confidence in the class decision. Additionally, the individual blocks are classified using class-specific residuals akin to SRC [27] and it is desirable to utilize this information for fusion.

Accordingly, we use a different approach to fuse individual decisions based intuitively on the maximum likelihood. Let $\hat{S}_i$ be the recovered sparse representation matrix of the block $Y_i$. The probability of $Y_i$ belonging to the $k$-th class is defined to be inversely proportional to the residual associated with the dictionary atoms in the $k$-th class:

$$p_i^k = P(\text{Class}(Y_i) = k) = \frac{1/R_i^k}{\sum_{k=1}^K (1/R_i^k)},$$

where $R_i^k$ is the residual associated with the $k$-th dictionary atom.
where \( R_k^i = \| Y_i - \tilde{D} S_k (\hat{S}_i) \|_2^2 \). The identity of the test image \( Y \) is then given by:

\[
\text{Class}(Y) = \arg \max_{k = 1, \ldots, K} \left( \prod_{i=1}^{B} p^k_i \right).
\]

The \( p^k_i \) may be interpreted intuitively in terms of the (frequentist) maximum likelihood. Essentially, the idea is that the class \( k \) that leads to the lowest reconstruction error for local block \( i \) corresponds to the largest “probability” or “likelihood” among all the \( p^k_i \). The subsequent fusion step merely seeks the class \( k \) that maximizes the product of all such “probabilities” from all the \( B \) local blocks.

E. The Role of Feature Extraction

SRC [27] was originally proposed as a method that views a vectorized image as a linear combination of vectorized training images. The richness (size of training set) of the training dictionary coupled with the low dimensions of the images leads to an over-complete dictionary. Expectedly, the curse of dimensionality strikes and renders the problem computationally intensive. Nonetheless, there is a precedent for the successful application of SRC directly on raw images in a variety of tasks such as face recognition [27], SAR automatic target recognition [29], hyperspectral image classification [48] and many more. Our approach is a continuation of efforts in this direction, the novelty being in the application to histopathology. The local version of our algorithm, LA-SHIRC, involves local image blocks of much smaller dimension compared to the entire image. This alleviates the issue of high dimensionality to an extent.

A more common way of handling high dimensionality is to work with image features instead of the images themselves. However, it is easier to interpret the linear representation model in SRC for images rather than on features. It is not apparent at first glance why a test feature should be a linear combination of training features. Also, one of the motivations for our work is the relative difficulty in designing histopathological image features due to the geometric richness of tissue images. The linear representation model alleviates this difficulty. That said, many recent approaches have employed SRC (or extensions) on image features (see [32], [37] for example). Our classifier can be similarly employed on image features that are chosen based on domain knowledge. Importantly, the structure of the coefficient matrix remains unchanged, and the same optimization problem is solved by Algorithm 1 in Appendix A. We present experimental evaluation of SHIRC and LA-SHIRC on features in Section IV.

IV. VALIDATION AND EXPERIMENTAL RESULTS

A. Experimental Set-Up: Image Data Sets

We compare the performance of SHIRC and LA-SHIRC against state-of-the-art alternatives for two challenging real-world histopathological image data sets.

ADL data set: These images are provided by pathologists at the Animal Diagnostics Lab, Pennsylvania State University. The tissue images have been acquired from three different mammalian organs - kidney, lung, and spleen. For each organ, images belonging to two categories - healthy or inflammatory - are provided. The H&E-stained tissues are scanned using a whole slide digital scanner at 40x optical magnification, to obtain digital images of pixel size \( 4000 \times 3000 \). All images are downsampled to \( 100 \times 75 \) pixels in an aliasing-free manner for the purpose of computational speed-up. The algorithm in fact works at any image resolution. Example images\(^2\) are shown in Fig. 6. There are a total of 120 images for each organ, of which 40 images are used for training and the rest for testing. The ground truth labels for healthy and inflammatory tissue

\(^2\)While the entire data set cannot be made publicly available, sample full-resolution images can be viewed at: http://signal.ee.psu.edu/histimg.html.
are assigned following manual detection and segmentation performed by ADL pathologists. We present classification results separately for each organ. In the experiments to follow, results using LA-SHIRC on the ADL data set are not reported since a single block (i.e. the entire image) was deemed by pathologists to have sufficient discriminative information. Essentially, the performance of LA-SHIRC is identical to SHIRC in this case.

It is worthwhile to briefly understand the biological mechanisms underlying the different conditions in these images. Inflammatory cell tissue in cattle is often a sign of a contagious disease, and its cause and duration are indicated by the presence of specific types of white blood cells. Inflammation due to allergic reactions, bacteria, or parasites is indicated by the presence of eosinophils. Acute infections are identified by the presence of neutrophils, while macrophages and lymphocytes indicate a chronic infection. In Fig. 6, we observe that a healthy lung is characterized by large clear openings of the alveoli, while in the inflamed lung, the alveoli are filled with bluish-purple inflammatory cells. Similar clusters of dark blue nuclei indicate the onset of inflammation in the other organs.

**IBL data set:** The second data set comprises images of human intraductal breast lesions [10]. The images belong to either of two well-defined categories: usual ductal hyperplasia (UDH) and ductal carcinoma in situ (DCIS). UDH is considered benign and the patients are advised follow-up check-ups, while DCIS is actionable and the patients require surgical intervention. Ground truth class labels for the images are assigned manually by the pathologists. A total of 40 patient cases - 20 well-defined DCIS and 20 UDH - are identified for experiments in the manner described in [10]. Each case contains a number of Regions of Interest (RoIs), and we have chosen a total of 120 images (RoIs), consisting of a randomly selected set of 60 images for training and the remaining 60 RoIs for test. Each RoI represents a full-size image for our experiments. Smaller local regions are chosen carefully within each such RoI for LA-SHIRC as described in III-C, using a classical morphology-based blob detection technique [12].

We compare the performance of SHIRC and LA-SHIRC with two competing approaches:

- **SVM:** this method combines state-of-the-art feature extraction and classification. We use the collection of features from WND-CHARM [19], [20] which is known to be a powerful toolkit of features for medical images. A support vector machine is used for decisions unlike weighted nearest neighbor in [19] to further enhance classification. We pick the most relevant features for histopathology [4], including but not limited to (color channel-wise) histogram information, image statistics, morphological features and wavelet coefficients from each color channel. The source code for WND-CHARM is made available by the National Institutes of Health online at: [http://ome.grc.nia.nih.gov/wnd-charm/](http://ome.grc.nia.nih.gov/wnd-charm/).

- **SRC:** the single channel sparse representation-based classification approach reviewed in Section II. Specifically, we employ SRC directly on the luminance channel (obtained as a function of the RGB channels) of the histopathological images, as proposed initially for face recognition and applied widely thereafter.

For results on the IBL data set, we also directly report corresponding numbers from [10] - the multiple-instance learning (MIL) algorithm - which is a full image analysis and classification system developed custom for the IBL data set.

In supervised classification, it is likely that some particularly well-chosen training sets can lead to high classification accuracy. In order to mitigate this issue of selection bias, we perform 10 different trials of each experiment. In each trial, we randomly select a set of training images – all results reported are the average of the classification accuracies from the individual trials.

**B. Validation of Central Idea: Overall Classification Accuracy**

First, we provide experimental validation of the central hypothesis of this paper: that exploiting color information in a principled manner through simultaneous sparsity models leads to better classification performance over existing techniques for histopathological image classification. To this end, we present overall classification accuracy for each of the three organs from the ADL data set, in the form of bar graphs in Fig. 10(a). SHIRC outperforms SVM and SRC in each of the three organs, thereby confirming the merits of our multi-channel simultaneous sparsity model in (13). The selection of application-specific features coupled with the inclusion of features from the RGB channels ensures that the SVM classifier performs competitively, particularly for the lung.

A similar experiment using the full-size images from the IBL data set illustrates the variability in histopathological imagery. Each image in the data set contains multiple cells at different spatial locations, as seen in Fig. 9. SHIRC is not designed to handle this practical challenge. The bar graph in Fig. 10(b) shows that the SVM classifier and the systemic MIL approach in [10] offer the best classification accuracy. This is not surprising because MIL [10] incorporates elaborate segmentation and pre-processing followed by feature extraction strategies customized to the acquired set of images. This experimental scenario occurs frequently enough in practice and serves as our motivation to develop LA-SHIRC.
C. Detailed Results: Confusion Matrices and ROC Curves

Next, we present a more elaborate interpretation of classification performance in the form of confusion matrices and ROC curves. Each row of a confusion matrix refers to the actual class identity of test images and each column indicates the classifier output.

Tables I-III show the mean confusion matrices for the ADL data set. In continuation of trends from Fig. 10(a), SHIRC offers the best disease detection accuracy - a quantitative metric of high relevance to pathologists - for each organ, while maintaining high classification accuracy for healthy images too. An interesting observation can be made from Table III. The SVM classifier reveals a tendency to classify the diseased tissue images much more accurately than the healthy tissues. In other words, there is a high false alarm rate (healthy image mistakenly classified as inflammatory) associated with the SVM classifier. SHIRC however offers a more consistent class-specific performance, resulting in the best overall performance. The corresponding results using LA-SHIRC are identical to SHIRC and hence not shown, since a single block (i.e. the entire image) was deemed by pathologists to have sufficient discriminative information.

Table IV shows the mean confusion matrix for the IBL data set. SHIRC provides an average classification accuracy of 66.09%, in comparison with about 87.9% using the MIL approach [10]. However, LA-SHIRC results in a significant improvement in performance, even better than the rates reported using SVM, MIL or SRC. The informed version of LA-SHIRC as described in Section III-C was implemented for the results in Table IV (and the ROC curves in Fig. 10) using - a pathologist recommended - 9 local blocks per image.

It is noteworthy that a pre-processing stage involving careful image segmentation is performed prior to feature extraction in MIL [4], implying that MIL is representative of state-of-the-art classification techniques using local image information.

Typically in medical image classification problems, pathologists desire algorithms that reduce the probability of miss (classifying diseased image as healthy) while also ensuring that the false alarm rate remains low. However, there is a trade-off between these two quantities, conveniently described using a receiver operating characteristic (ROC) curve. Fig. 11 shows the ROC curves for the ADL and IBL data sets. The lowest curve (closest to the origin) has the best overall performance and the optimal operating point minimizes the sum of the miss and false alarm probabilities. In Figs. 11(a)-(c), SHIRC offers the best trade-off. In Fig. 11(d), the LA-SHIRC outperforms SVM, and both methods are much better than SRC and SHIRC.

Remarks: Note that ROCs for MIL [10] could not be reported because the image analysis and classification system in [10] has a variety of pre-processing, segmentation and other image processing and classification steps which makes exact reproduction impossible in the absence of publicly available code.
Also note that in Fig. 11(a)-(c), the ROC curve for LA-SHIRC has not been reported because SHIRC and LA-SHIRC are identical; the pathologists recommended using a single block, i.e. the entire image, that captures discriminative information.

Depending on the inherent degree of difficulty in classifying a particular image set and the severity of the penalty for misclassifying a diseased image\(^3\), a pathologist can choose an acceptable probability of miss and corresponding false alarm rate for each method. Table V shows that for each organ in the ADL data set, a higher false alarm must be tolerated with the SVM method, compared to SRC and SHIRC, in order to maintain a fixed rate of miss. For the IBL data set, the LA-SHIRC incurs the lowest false alarm rate to achieve a miss rate of 10%.

\(^3\)For example, DCIS requires immediate surgical attention, while a mild viral infection may only prolong for a few more days if not diagnosed early.

### D. Performance vs. Size of Training Image Set

This experiment offers new insight into the practical performance of our algorithms. Real-world classification tasks often suffer from the lack of availability of large training sets. We present a comparison of overall classification accuracy as a function of the training set size for the different methods.

In Fig. 12(a), overall classification accuracy is reported for ADL data set (kidney) corresponding to five different training scenarios. The case of 20 training images per class pertains to low training, while 40 training images per class represents adequate training for this data set. As before, comparison is made against the single channel SRC and state-of-the-art feature extraction plus SVM classifier. Unsurprisingly, all three methods suffer in performance as training is reduced.

Fig. 12(b) reports analogous results for the IBL data set. Here the regime of low training is defined by 20 images per class, while the adequate training scenario is captured by 60 images per class. Analyzing the results in Fig. 12(b) for the IBL data set, a more interesting trend reveals itself. As discussed before in Section III-C, LA-SHIRC can lead to richer dictionaries made out of local image blocks even as the number of training images is not increased. This allows LA-SHIRC to perform extremely well even under low training - offering about 90% accuracy - as is evident from Fig. 12(b). This benefit however comes at the cost of increased computational complexity at the time of inference because the dictionary size (number of columns) is significantly increased.
E. Performance vs. Number of Local Blocks

In Section III-D, we propose two different strategies to select the local blocks that capture discriminative information. For each strategy - informed and blind selection - we select different numbers of local blocks from the test images, classify them separately according to the pipeline in Fig. 7, and fuse the decisions using the technique discussed in Section III-D.

For the results presented so far, we have used 9 well-selected local blocks for the IBL data set. Fig. 13 shows the variation in classification performance of LA-SHIRC as the number of local blocks is increased, both for informed and blind selection. We use 20 training images per class, which is in fact representative of low training as discussed in Section IV-D. Two significant trends emerge from this figure. First, unsurprisingly informed block selection leads to better performance even with a smaller number of blocks. The performance of the blind selection strategy is poor when a small number of blocks is chosen. However, when sufficiently large number of blocks are included for decision-making, its performance is very similar to that of informed selection. The benefit of blind selection is that it can potentially circumvent the need for elaborate segmentation/block detection techniques albeit at the cost of additional computation. Second, we observe that in the informed selection, the performance only improves slightly as the number of local blocks is increased beyond a certain number (in this case 9). This is also intuitively satisfying because informed selection is based on explicit pathologist input in terms of both size and number of blocks necessary for adequately good discrimination.

F. Evaluation: Classification of Image Features

In this experiment, we evaluate SRC and SHIRC on a good choice of histopathological image features obtained from [19]. The results are shown in Fig. 14 and Table VI. In comparison with Table II, it can be seen that the classification accuracy for SHIRC with features is in fact inferior to the classification accuracy on the raw images. This shows that feature design for histopathology is a challenging task. It also confirms that our proposed SHIRC alleviates this burden on feature selection effectively.

G. Runtime Comparison

It is well-known that SRC (and its extensions) have a (computationally) trivial training stage - stacking the vectors into a dictionary - unlike the SVM, which involves the solution of a quadratic program to identify the support vectors. Significantly, this training stage is entirely performed
We only report the time taken to classify amounts to evaluating a linear combination of a small set of relaxations. The classification for SVM, on the other hand, (3) in realtime via greedy approaches [40], [41] or convex in the test stage, SRC solves the optimization problem in Eq. offline. When a previously unseen vector is being classified in the test stage, SRC solves the optimization problem in Eq. (3) in realtime via greedy approaches [40], [41] or convex relaxations. The classification for SVM, on the other hand, amounts to evaluating a linear combination of a small set of inner products (with the support vectors). In our comparison, we only report the time taken to classify test vectors for SRC, SHIRC and LA-SHIRC. The results are obtained by running MATLAB (version 2012a) on a 64-bit Windows 7 system equipped with Intel(R) Core i7-2600 3.4 GHz processor and 8 GB RAM.

SRC (using the greedy basis pursuit) takes about 0.13 seconds on average, while SHIRC takes about 0.55 seconds to classify a single test image. Assuming that the classification of each local block is performed in sequence on a single core, it is not surprising that the performance of LA-SHIRC increases linearly as the number of local blocks used. The time taken to classify a single local block is about 0.35 seconds. This is lesser than the SHIRC runtime since the dimension of the local block is typically much smaller in comparison to the size of the entire image. For three and nine local blocks, the LA-SHIRC runtime is about 0.48 seconds and 1.09 seconds respectively.

There is a trade-off between accuracy and computational requirements. The upside of LA-SHIRC is the robustness in classification performance, even with a small number of training samples, which is more essential than fast realtime operation in histopathological imaging applications.

### H. Outlier Rejection

Sometimes in histopathological image classification, an image may show up from a hitherto unseen class. Thus it is desirable for a classification algorithm to exhibit the property of outlier rejection. The class decision fusion in LA-SHIRC, shown in (18), naturally leads to a metric to identify outliers. The overall class decision is the class that maximizes the product of the individual class “likelihoods.” We flag a test image as an outlier if this product is less than a well-chosen threshold $\tau$.

A more effective way of checking for outliers is inspired by the sparsity concentration index (SCI) in SRC [27]. As a natural extension of this metric for the joint sparsity problem, we use a joint version of SCI as follows:

$$JSCI(S') := \frac{K \max_k \| \delta_k(S')^\top \|_{row,0} / \| S' \|_{row,0} - 1}{K - 1},$$  \hspace{1cm} (19)

where $K$ is the number of classes. This index takes on values in the range $[0, 1]$. A value of $JSCI(S')$ close to 1 indicates that the test sample belongs to one of the trained classes, whereas a value close to 0 indicates an outlier.

In this experiment, we build our dictionary from two classes of liver tissue images - healthy and inflammatory - and test for outlier rejection using a third class (necrotic). These images have been acquired as part of the ADL data set. We trained the dictionary with 20 images per class and tested over 40 outlier test images. We chose a threshold of 0.13 which is obtained experimentally. Then if $JSCI$ value is greater than this threshold it means the test sample is indeed from one of the trained classes, otherwise it means an outlier. Using the JSCI results in outlier rejection rates of about 68% in SRC and 75% in SHIRC. Similar trends can be seen when we use a threshold on the product of probabilities in (18).

### I. Summary of Results and Reproducibility

Here, we summarize the key messages from the various experiments described in Section IV-B-IV-F. Our first aim is to validate our central hypothesis, that a multi-channel simultaneous sparsity model can be used to represent and classify color histopathological images. Accordingly, we demonstrate the improvements offered by our proposed approach (SHIRC) over state-of-the-art alternatives. While the expected classification trends are seen for the ADL data set, the IBL data set presents a challenging scenario where the SHIRC method, if applied directly, could lead to worse performance than a traditional SVM classifier. We then show more elaborate experimental results, in the form of confusion matrices and ROC curves, to demonstrate that LA-SHIRC, the locally adaptive variant of our SHIRC, indeed outperforms all other techniques by creating richer dictionaries that are synthesized by combining local image blocks. A new experimental insight is gained in that LA-SHIRC offers high classification accuracy even with a limited number of training images. Next, we show the performance of LA-SHIRC as a function of the number of local blocks to highlight the complementary benefits of the local block selection strategies. Finally, we discuss the following issues of practical interest: the role of image features
for classification, a comparison of algorithm runtimes and the ability of the proposed algorithms to reject images from unseen classes.

In order to facilitate the use of our proposed algorithms for medical image classification and other multi-variate/multi-task classification problems, the MATLAB code for the algorithms and all experiments described here is posted online at: http://signal.ee.psu.edu/histimg.html.

V. DISCUSSION AND CONCLUSION

A. Summary

In this paper, we have proposed a simultaneous sparsity model for histopathological image representation and classification. The central idea of our approach is to exploit the correlations among the red, green and blue channels of the color images in a sparse linear model setting with attendant color channel constraints. We formulate and solve a new sparsity-based optimization problem. We also introduce a robust locally adaptive version of the simultaneous sparsity model to address the issue of correspondence of local image objects located at different spatial locations. This modification results in benefits that have significant practical relevance: we demonstrate that the sparsity model for classification can work even under limited training if local blocks are chosen carefully.

B. Future Work

A natural future direction for this work is to deploy SHIRC (and LA-SHIRC) widely as an important diagnostic tool in existing histopathological image analysis systems such as [10]. Particular problems such as grading in histopathology [9],[17] and multimodal fusion for disease diagnosis [3] can be investigated using our proposed techniques.

Sparse representation-based image classification is an area of ongoing research interest, and here we identify some connections to our work in published literature. Our framework can be generalized for any multi-variate/multi-task classification problem [49] by simply including training from those tasks as new sub-dictionaries. Recent work in multi-task classification has explored the idea of sparse models on image features [37]. Admittedly, the sparse linear model may not be justifiable for all types of histopathological classification problems. One way of incorporating non-linear sparse models is to consider the sparse model in a feature space induced by a kernel function [48]. Recent work has focused attention on solving the costly sparsity optimization problem more effectively [50]–[52]. We believe a deeper investigation towards efficient solutions to our modified optimization problem is a worthwhile research pursuit.

ACKNOWLEDGMENT

The authors thank Prof. Murat Dundar of IUPUI for providing the breast lesion images used in our experiments.

APPENDIX A

A GREEDY PURSUIT APPROACH TO MULTI-TASK CLASSIFICATION

Notation: Let $y_i \in \mathbb{R}^n, i = 1, \ldots, T$ be $T$ different representations of the same physical event, which is to be classified into one of $K$ different classes. Let $Y := [y_1 \ldots y_T] \in \mathbb{R}^{n \times T}$. Assuming $n$ training samples/events in total, we design $T$ dictionaries $D_i \in \mathbb{R}^{m \times n}, i = 1, \ldots, T$, corresponding to the $T$ representations. We define a new composite dictionary $D := [D_1 \ldots D_T] \in \mathbb{R}^{m \times nT}$. Further, each dictionary $D_i$ is represented as the concatenation of the sub-dictionaries from all classes corresponding to the $i$-th representation of the event:

$$ D_i := [D_i^1 \ D_i^2 \ \ldots \ D_i^K], \quad (20) $$

where $D_i^j$ represents the collection of training samples for representation $i$ that belong to the $j$-th class. So, we have:

$$ D := [D_1 \ldots D_T] = [D_1^1 \ D_1^2 \ \ldots \ D_1^K \ D_2^1 \ D_2^2 \ \ldots \ D_2^K \ \ldots \ D_T^1 \ D_T^2 \ \ldots \ D_T^K]. \quad (21) $$

A test event $Y$ can now be represented as a linear combination of training samples as follows:

$$ Y = [y_1 \ldots y_T] = DS $$

$$ = [D_1^1 \ D_1^2 \ \ldots \ D_1^K \ D_2^1 \ D_2^2 \ \ldots \ D_2^K \ \ldots \ D_T^1 \ D_T^2 \ \ldots \ D_T^K] [\alpha_1 \ \ldots \ \alpha_T], $$

where the coefficient vectors $\alpha_i \in \mathbb{R}^{n_T}, i = 1, \ldots, T$, and $S = [\alpha_1 \ \ldots \ \alpha_T] \in \mathbb{R}^{n_T \times T}$.

Since $S$ obeys column correspondence, we introduce a new matrix $S' \in \mathbb{R}^{n \times T}$ as the transformation of $S$ with the zero coefficients removed,

$$ S' = \begin{bmatrix} \alpha_1^1 & \ldots & \alpha_1^j & \ldots & \alpha_1^K \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \alpha_T^1 & \ldots & \alpha_T^j & \ldots & \alpha_T^K \end{bmatrix}, $$

where $\alpha_i^j$ refers to the sub-vector extracted from $\alpha_i$ that corresponds to coefficients from the $j$-th class. Note that, in the $i$-th column of $S'$, only the coefficients corresponding to $D_i$ are retained (for $i = 1, \ldots, T$).

We can now apply row-sparsity constraints similar to the approach in [43]. Our modified optimization problem becomes:

$$ \hat{S}' = \arg\min_S ||S'||_{\text{row},0} \quad \text{subject to} \quad ||Y - DS||_F \leq \epsilon, \quad (22) $$

for some tolerance $\epsilon > 0$. We minimize the number of non-zero rows, while the constraint guarantees a good approximation.

The matrix $S$ can be transformed into $S'$ by introducing matrices $H \in \mathbb{R}^{nT \times T}$ and $J \in \mathbb{R}^{n \times nT}$,

$$ H = \text{diag} [1 \ 1 \ \ldots \ 1], \ J = [I_n \ I_n \ \ldots \ I_n], $$

where $1 \in \mathbb{R}^n$ is the vector of all ones, and $I_n$ denotes the $n$-dimensional identity matrix. Finally, we obtain $S' = J (H \circ S)$, where $\circ$ denotes the Hadamard product, $(H \circ S)_i \triangleq h_{ij}s_{ij}$ for all $i, j$. Eq. (22) represents a hard optimization problem due to presence of the non-invertible transformation from $S$ to $S'$. We bypass this difficulty by proposing a modified version of the SOMP algorithm for the multi-task multivariate case.

Recall that the original SOMP algorithm gives $K$ distinct atoms (assuming $K$ iterations) from a dictionary $D$ that best
Algorithm 1: SOMP for multi-task multivariate sparse representation-based classification

Input: Dictionary $D$, signal matrix $Y$, 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:
$$\lambda_{i,k} = \arg \max_{j=1,\ldots,n} \sum_{q=1}^{T} w_q \left\| R_{k-1}^{j} d_{q,j} \right\|_p, p \geq 1$$

2. Update the index set $\Lambda_{i,k} = \Lambda_{i,k-1} \cup \{\lambda_{i,k}\}$, $i = 1, \ldots, T$

3. Compute the orthogonal projector $p_{i,k} = \left( D_{\Lambda_i,k}^{T} D_{\Lambda_i,k} \right)^{-1} D_{\Lambda_i,k}^{T} y_i$, for $i = 1, \ldots, T$, where $D_{\Lambda_i,k} \in \mathbb{R}^{n \times k}$ consists of the $k$ atoms in $D_i$ indexed in $\Lambda_i,k$

4. Update the residual matrix $R_k = Y - \left[ D_{\Lambda_1,k} p_{1,k} \cdots D_{\Lambda_T,k} p_{T,k} \right]$

5. Increment $k$: $k \leftarrow k + 1$

end while

Output: Index set $\Lambda_i = \Lambda_{i,k}, i = 1, \ldots, T$; sparse representation $\hat{S}$ whose non-zero rows indexed for each representation by $\Lambda_i$, $i = 1, \ldots, T$, are the $K$ rows of the matrix $\left( D_{\Lambda_1,k}^{T} D_{\Lambda_1,k} \right)^{-1} D_{\Lambda_1,k}^{T} Y$.

represent the data matrix $Y$. In every iteration $k$, SOMP measures the residual for each atom in $D$ and creates an orthogonal projection with maximal correlation. Extending this to the multi-task setting, for every representation $i, i = 1, \ldots, T$, we can identify the index set that gives the highest correlation with the residual at the $k$-th iteration as follows:

$$\lambda_{i,k} = \arg \max_{j=1,\ldots,n} \sum_{q=1}^{T} w_q \left\| R_{k-1}^{j} d_{q,j} \right\|_p, p \geq 1,$$

where $w_q$ denotes the weight (confidence) assigned to the q-th representation, $d_{q,j}$ represents the j-th column of $D_q$, $q = 1, \ldots, T$, and the superscript $(\cdot)^T$ indicates the matrix transpose operator. After finding $\lambda_{i,k}$, we modify the index set to:

$$\Lambda_{i,k} = \Lambda_{i,k-1} \cup \lambda_{i,k}, i = 1, \ldots, T.$$

Thus, by finding the index set for the $T$ distinct representations, we can create an orthogonal projection with each of the atoms in their corresponding representations. The algorithm is summarized in Algorithm 1.

REFERENCES


