Morphologically Decoupled Multi-Scale Sparse Representation for Hyperspectral Image Analysis

Saurabh Prasad, *Senior Member, IEEE*, Minshan Cui, Demetrio Labate, *Member, IEEE* Yuhang Zhang

Abstract

Hyperspectral imagery has emerged as a popular sensing modality for a variety of applications, and sparsity based methods were shown to be very effective to deal with challenges coming from high dimensionality in most hyperspectral classification problems. In this work, we challenge the conventional approach to hyperspectral classification, that typically builds sparsity-based classifiers directly on spectral reflectance features or features derived directly from the data. We assert that hyperspectral image processing can benefit very significantly by decoupling data into geometrically distinct components since the resulting decoupled components are much more suitable for sparse representation based classifiers. Specifically, we apply morphological separation to decouple data into texture and cartoon-like components, which are sparsely represented using local discrete cosine bases and multiscale shearlets, respectively. In addition to providing sparser representation, this approach allows us to take advantage of the invariance properties of each basis within each geometrically distinct component of the data. Experimental results using real-world hyperspectral image datasets demonstrate the efficacy of the proposed framework for classifying multi-channel imagery under a variety of adverse conditions — in particular, small training sample size, additive noise, and rotational variabilities between training and test samples.

Index Terms

This material is based upon work supported in part by the 2013 NASA New Investigator (Career) award (project number: NNX14AI47G) and NSF-DMS (project number: 1320910).

S. Prasad, M. Cui, Y. Zhang are with the Hyperspectral Image Analysis Laboratory in the Department of Electrical and Computer Engineering, University of Houston. (e-mail: saurabh.prasad@ieee.org). D. Labate is with the Department of Mathematics at the University of Houston.

Hyperspectral data, multi-resolution analysis, sparse representation, image analysis.

I. Introduction

Hyperspectral imaging modalities have been extremely successful in a wide variety of applications, including remote sensing for ground cover analysis, terrestrial/ground based imaging for scene understanding, microscopy and other laboratory imaging for biomedical applications [1]–[6]. However, while remarkable technological advances in optics, electronics and integration of imaging systems with a variety of platforms have further increased the popularity and adoption of hyperspectral imaging modalities in recent years, there is a critical need to develop improved image analysis frameworks tailored to such data. Popular applications of hyperspectral images (HSI) include classification, spectral unmixing, change and anomaly detection [6]–[9]. Since the approach proposed in this paper is for robust HSI classification, the following discussion will be restricted to this topic and related work, although the ideas presented in this paper are applicable to applications such as anomaly and change detection and target recognition as well.

Traditional approaches to hyperspectral image (HSI) analysis, particularly classification, entail the following flow: Analysis typically begins with feature extraction and feature reduction (i.e., extracting pertinent spectral and spatial information, finding lower dimensional subspaces that preserve underlying information etc.), followed by design and optimization of classifiers that operate on the resulting feature space. Multispectral and hyperspectral sensors, when used for image analysis tasks typically result in very high dimensional feature spaces — posing unique challenges, such as burdening transmission and storage systems, reducing generalization capability of traditional Bayesian classifiers, etc. [2]–[4], [7], [10], [11]. Classical transform-based feature reduction approaches such as Principal Components Analysis (PCA), Independent Component Analysis (ICA), Fisher's Linear Discriminant Analysis - (LDA) and their many variants have been extensively developed and widely studied for hyperspectral image analysis tasks. More recently, supervised subspace learning approaches that exploit (and often preserve) the underlying structure of the data (e.g. data that resides on a manifold) have shown great promise for hyperspectral image analysis [12], [13]. These approaches have primarily alleviated the over-dimensionality problem posed hyperspectral data, better conditioning feature spaces for interpretation (classification, spectral unmixing, sub-pixel and pixel-level target or anomaly detection, change detection etc.). In separate developments, multi-resolution analysis based representations including wavelets, curvelets etc. have been utilized for hyperspectral data processing and analysis, facilitating improved compression, noise robust classification, denoising etc.

Sparse representations have emerged as a promising tool for a range of applications, including compressed sensing, signal denoising, and more recently, classification. In such representations, most or all of the information of an unknown signal can be linearly represented by a small number of atoms in a "dictionary". Based on this theory, a sparse representation classifier (SRC) was developed for robust face recognition, and was later adapted for other applications, including hyperspectral image classification. The central idea in SRC and its variants is to represent a testing sample (e.g. a pixel in a hyperspectral image) as a linear combination of all available training samples (which form an over-complete dictionary) [14]–[20] — most of the nonzero or large value entries in the recovered coefficients are expected to correspond to training samples having the same class membership as the testing sample. The assumption of such an approach is that the testing sample approximately lies in the linear span of the training samples from the same class. We note that virtue of their design, such approaches are generally robust to small training sample sizes, even when the dimensionality of the input space is large (e.g. with hyperspectral imagery).

Related to sparsity is the notion of *geometric separation*, according to which sparse representations can be used to exactly separate data consisting of geometrically distinct components (e.g., texture, smooth regions, egdes) provided one selects appropriately sparse representation systems which are also mutually incoherent [21], [22]. One classical manifestation of this idea is the Morphological Component Analysis (MCA), showing that images containing different morphologies can be broken into separate morphological components using an appropriate dictionary amalgamating multiple bases. This idea was successfully applied to problems of image denoising and restoration [23]–[25].

Inspired by these ideas, in this paper *we introduce a novel framework of image classification for robust image understanding based on a morphologically decoupled sparse representation and customized to high dimensional hyperspectral imagery (we validate with hyperspectral imagery, but the ideas can be applied to any multi-channel imagery)*. The proposed approach utilizes a bank of sparse representation classifiers operating on a sequence of subspaces generated via MCA, with each classifier optimized to a dictionary that provides optimally sparse representation on a specific subspace. The core idea of our approach is that hyperspectral imaging data (be it remote sensing or images of natural scenes or of biological samples) can be modeled as superposition of multiple geometrically distinct components, e.g., texture-like and a cartoon-like components. By building a combined dictionary consisting of sub dictionaries that are optimally sparse in each distinct image component, we obtain a data representation *adapted to the geometry of each image component*. We contend that the classifiers resulting from this approach are *morphologically optimal* in the sense that they use the sparsest representation for each image component. An additional benefit of this approach is that we can exploit the special 'geometric' properties of these

morphologically-adapted data representations to derive classifiers endowed with rotational invariance and noise robustness.

The remainder of this paper is organized as follows. In section II, we provide background on sparse shearlet representations, the notion of combined dictionaries and multi-task joint sparse representation based classification. In section III, we describe the proposed classification approach and provide insights specific to specific benefits with this framework — orientation invariance and noise robustness. In section IV, we validate the proposed approach by applying it to two real-world hyperspectral datasets. We demonstrate the efficacy of the proposed approach, and quantify the associated benefits of orientation invariance and noise robustness. Concluding remarks are provided in section V.

II. Background and Related Work

A. Sparse shearlet Representation

The shearlet representation emerged during the last decade as a powerful refinement of conventional wavelets and other classical multiscale representations [26], [27]. Similar to curvelets [28], shearlets are well-localized waveforms defined not only over a range of scales and locations, like wavelets, but also over multiple orientations and with highly anisotropic shapes so that they are especially efficient to capture edges and the other relevant geometric features in images.

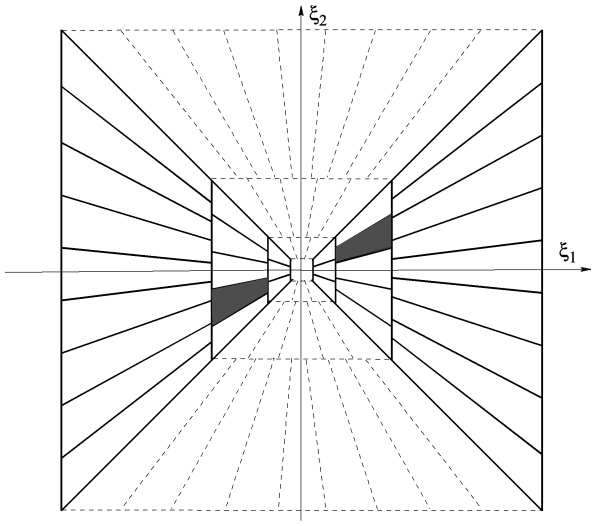


Fig. 1: **Tiling of the Fourier plane associated with shearlets ($n = 2$). The Fourier support $\Sigma_{j,\ell}$ of a typical shearlet element is shown as a solid grey region. The horizontal and vertical cones are partitioned into directional subbands using solid lines and dashed lines, respectively.**

Roughly speaking, in dimension $n = 2$, shearlets are generated by the action of anisotropic dilations and shear transformations on an appropriate set of generators $\psi^{(\nu)} \in L^2(\mathbb{R}^2)$, that is,

$$\psi^{(\nu)}_{j,\ell,k}(x) = 2^{3j/2}\psi^{(\nu)}(B^{\ell}_{\nu}A^{j}_{\nu}x - k), \tag{1}$$

for $j \geq 0$, $-2^j \leq \ell \leq 2^j$, $k \in \mathbb{Z}^2$, $\nu = 1, 2$, where $A_1 = \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix}$, $A_2 = \begin{pmatrix} 2 & 0 \\ 0 & 4 \end{pmatrix}$ are the *anisotropic dilation matrices* and $B_1 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$, $B_2 = B_1^t$ are the *shear matrices*. The indices j, ℓ, k are associated with a range of scales, orientations and locations, respectively.

Shearlet properties are better illustrated by the precise shearlet construction in the Fourier domain [29].

Let $\phi \in C^{\infty}([0, 1])$ be a 'bump' function with $\operatorname{supp}\phi \subset [-\frac{1}{8}, \frac{1}{8}]$ and $\phi = 1$ on $[-\frac{1}{16}, \frac{1}{16}]$. For $\xi = (\xi_1, \xi_2) \in \mathbb{R}^2$, let $\Phi(\xi) = \Phi(\xi_1, \xi_2) = \phi(\xi_1)\,\phi(\xi_2)$ and define

$$W(\xi) = W(\xi_1, \xi_2) = \sqrt{\Phi^2(2^{-2}\xi_1, 2^{-2}\xi_2) - \Phi^2(\xi_1, \xi_2)}.$$

The functions $W_j^2 = W^2(2^{-2j}\cdot)$, $j \geq 0$, have support inside the Cartesian coronae

$$C_j = [-2^{2j-1}, 2^{2j-1}]^2 \setminus [-2^{2j-4}, 2^{2j-4}]^2$$

and produce a smooth tiling of the frequency plane:

$$\Phi^2(\xi_1, \xi_2) + \sum_{j \geq 0} W^2(2^{-2j}\xi_1, 2^{-2j}\xi_2) = 1 \quad \text{for } (\xi_1, \xi_2) \in \mathbb{R}^2.$$

To obtain an angular partition, let $V \in C^{\infty}(\mathbb{R})$ so that $\operatorname{supp} V \subset [-1, 1]$, $V(0) = 1$,

$$|V(u-1)|^2 + |V(u)|^2 + |V(u+1)|^2 = 1 \quad \text{for } |u| \leq 1.$$

Hence, the 'fine-scale' shearlets are the functions

$$\begin{aligned} \hat{\psi}^{(\nu)}_{j,\ell,k}(\xi) = 2^{-3j/2}\, & W(2^{-j}\xi)\, F_{\nu}(\xi A_{\nu}^{-j} B_{\nu}^{-\ell}) \\ & \times e^{2\pi i \xi A_{\nu}^{-j} B_{\nu}^{-\ell} k}, \end{aligned} \tag{2}$$

where $F_1(\xi_1, \xi_2) = V(\frac{\xi_2}{\xi_1})$ and $F_2(\xi_1, \xi_2) = V(\frac{\xi_1}{\xi_2})$ and the matrices A_ν, B_ν are as above. As shown in [29], functions (2) can be (essentially) written in space-domain as (1).

We remark that the functions $\hat{\psi}^{(1)}_{j,\ell,k}$ can be written as

$$\hat{\psi}^{(1)}_{j,\ell,k}(\xi) = 2^{-2j}\, W(2^{-2j}\xi)\, V\Big(2^j \frac{\xi_2}{\xi_1} - \ell\Big)\, e^{2\pi i \xi A_1^{-j} B_1^{-\ell} k},$$

showing that their supports are contained inside the trapezoidal regions

$$\Sigma_{j,\ell} = \{(\xi_1, \xi_2) : 2^{2j-4} < |\xi_1| < 2^{2j-1}, |\tfrac{\xi_2}{\xi_1} - \ell 2^{-j}| \leq 2^{-j}\}$$

within the *horizontal cone* $|\xi_2| \leq |\xi_1|$ of the Fourier plane. Similar properties hold for the functions $\hat{\psi}^{(2)}_{j,\ell,k}$, whose supports are contained within the *vertical cone* $|\xi_2| > |\xi_1|$ of the Fourier plane. The tiling of the Fourier plane associated with the shearlet construction is shown in Fig. 1.

A smooth Parseval frame for $L^2(\mathbb{R}^2)$ is obtained by combining the 'fine-scale' shearlets together with a coarse scale system, associated with the low frequency region.[1] That is, we define a *shearlet system for* $L^2(\mathbb{R}^2)$ as

$$\left\{\widetilde{\psi}_{-1,k} : k \in \mathbb{Z}^2\right\} \cup \left\{\widetilde{\psi}_{j,\ell,k,\nu} : j \geq 0, |\ell| < 2^j, k \in \mathbb{Z}^2, \nu = 1, 2\right\}$$

where $\widetilde{\psi}_{-1,k} = \check{\Phi}(\cdot - k)$ and $\widetilde{\psi}_{j,\ell,k,\nu} = \psi^{(\nu)}_{j,\ell,k}$. With compact notation, we denote this system as

$$\{\widetilde{\psi}_\mu,\ \mu \in M\}, \tag{3}$$

where $M = M_C \cup M_F$ are the indices associated with coarse-scale and fine-scale shearlets, respectively; that is, $M_C = \{(j,k) : j = -1, k \in \mathbb{Z}^2\}$, $M_F = \{(j,\ell,k,\nu) : j \geq 0, |\ell| < 2^j, k \in \mathbb{Z}^2, \nu = 1, 2)\}$. We have the following result from [29]:

Theorem 1: The system of shearlets (3) *is a Parseval frame for* $L^2(\mathbb{R}^2)$. *That is, for any* $f \in L^2(\mathbb{R}^2)$, *we have the reproducing formula*

$$f = \sum_{\mu \in M} \langle f, \widetilde{\psi}_\mu \rangle \, \widetilde{\psi}_\mu,$$

with convergence in the L^2*-norm. All elements* $\{\widetilde{\psi}_\mu,\ \mu \in M\}$ *are* C^∞ *and compactly supported in the Fourier domain.*

By combining multiscale analysis and high directional sensitivity, shearlets can provide highly sparse representations for a large class of multidimensional data, outperforming conventional multiscale representations. In particular, for *cartoon-like* functions[2], a simplified model of images with edges, they are (nearly) optimally sparse [30].

Theorem 2: Let $f \in E^2$, *the class of carton-like functions in* $\mathbb{R}^2$, *and* f_N *be its* N*-term approximation obtained by taking the* N *largest coefficients in the shearlet representation of* f. *Then:*

$$\|f - f_N\|_2^2 \leq C\, N^{-2} (\log N)^3.$$

Ignoring the log factor, this result yields the optimal decay rate (no other basis or frame can achieve faster decay rate than N^{-2}) and it outperforms, in particular, wavelet approximations whose error rate is only of order N^{-1}. We remark that curvelets achieve the same type of approximation rate [28].

[1]To ensure that all elements of this combined shearlet system are C^∞_c in the Fourier domain, the elements whose supports overlap the boundaries of the cone regions in the frequency domain are slightly modified [29].

[2]Roughly speaking, this is the class of functions that are C^2 regular away from C^2 edges [30].

B. Combined Dictionaries and Morphological Separation

While shearlets and curvelets provide optimally sparse approximations for cartoon-like images, they may be not as efficient when dealing with images of natural scenes acquired on the ground or through remote sensing platforms.

Data found in many applications (including HSI) are often complex and there is no single representation system that can optimally approximate all the features of interest. One powerful strategy to address this situation is based on the principles of *morphological component analysis* (MCA), whose central idea (pioneered in [25], [31]) is to use multiple dictionaries to break up an image into its elementary geometric constituents. For this strategy to be effective, the various dictionaries are chosen to be mutually incoherent. That is, each dictionary leads to sparse representations for its intended image type, while yielding nonsparse representations on the other image type.

In this paper, we consider an approach that applies the MCA framework to sparsely represent an image x with respect to a combined dictionary. The method we present here is adapted from an application to data denoising in [24]. We assume that the discrete image x is a superposition of two geometrically distinct components

$$x = x_p + x_t, \tag{4}$$

where x_p is the piecewise smooth component of the data and x_t its textured component. To represent x, we use a dictionary $\mathcal{D}$ built by amalgamating two subdictionaries $\mathcal{D}_p$ and $\mathcal{D}_t$ that are 'incoherent'. That is, each component of x has a sparse representation in one subdictionary but its representation in the other subdictionaries is not sparse. In particular, for the subdictionary associated with texture component of the data we choose a local discrete cosine dictionary, which is sparse for locally periodic patterns. For the piecewise smooth component of the data, we choose a shearlet dictionary, which is known to be sparse for this type of data. The incoherence of the two dictionaries has been verified heuristically in [31] (using DCT and curvelet dictionaries) and more recently and rigorously in [21], [22]. Clearly, when we write x with respect to the overcomplete dictionary $\mathcal{D}$, as

$$x = \mathcal{D}\alpha = \sum_{k=1}^{K} \alpha_k \, d_k,$$

there are many possible expansions. In order to minimize the number of non-negligible coefficients, we can set up the minimization problem [32]

$$\hat{\alpha} = \min \|\alpha\|_1 \quad \text{subject to } \|x - \mathcal{D}\alpha\|_2 \leq \sigma, \tag{5}$$

where σ is the standard deviation of the noise, and so compute $\hat{x} = \mathcal{D}\hat{\alpha}$. Note that, for an appropriate parameter η, the solution of (5) is exactly the solution of the unconstrained optimization problem[3]

$$\min_{\alpha} \eta\|\alpha\|_1 + \frac{1}{2}\|x - \mathcal{D}\alpha\|_2^2. \tag{6}$$

To better exploit the geometric structure of the data, we can look for an expansion that takes advantage of the sparsity of the two subdictionaries. This is achieved by setting the minimization problem:

$$\{\hat{\alpha}_t, \hat{\alpha}_p\} = \min_{\alpha_t, \alpha_p} \eta\left(\|\alpha_t\|_1 + \|\alpha_p\|_1\right) + \frac{1}{2}\|x - \mathcal{D}_t\,\alpha_t - \mathcal{D}_p\,\alpha_p\|_2^2, \tag{7}$$

where $\mathcal{D}_p$, $\mathcal{D}_t$ are the dictionary associated with the piecewise smooth component and textured component of the data, respectively. The restored value of x is then found by adding together the two components obtained as $\hat{x}_p = \mathcal{D}_p\hat{\alpha}_p$ and $\hat{x}_t = \mathcal{D}_t\hat{\alpha}_t$. Note that, since our subdictionaries are tight frames, then $\mathcal{D}_p$ is the Moore-Penrose pseudo inverse of the analysis operator $\mathcal{W}_p$ associated with piecewise smooth data, i.e. $\mathcal{D}_p = \mathcal{W}_p^{\dagger}$ and, similarly, $\mathcal{D}_t$ is the Moore-Penrose pseudo inverse of the analysis operator $\mathcal{W}_t$ associated with texture data, i.e., $\mathcal{D}_t = \mathcal{W}_t^{\dagger}$.

Following the approach in [24], rather than using a sparsity-based *synthesis model* as in (7), we adopt a sparsity-based *analysis model* leading to the minimization problem

$$\{\hat{x}_p, \hat{x}_t\} = \underset{x_p, x_t}{\operatorname{argmin}}\, \eta\|\mathcal{W}_p x_p\|_1 + \eta\|\mathcal{W}_t x_t\|_1 + \tfrac{1}{2}\|x - x_p - x_t\|_2^2 \tag{8}$$

While in the synthesis formulation signals are modeled as sparse linear combinations of dictionary atoms, the analysis formulation emphasizes the zeros in the analysis side (rather than the non-zeros), leading to better performance. Another advantages of using the formulation (8) rather than (7) is that it requires searching lower dimensional vectors rather than longer dimensional representation coefficient vectors. To further improve the performance, we also included a total variation regularization term, which is effective at reducing possible ringing artifacts near the edges [31]. Thus, we finally have the optimization problem:

$$\begin{aligned}\min_{x_p, x_t} &\eta\|\mathcal{W}_p\,x_p\|_1 + \eta\|\mathcal{W}_t\,x_t\|_1 + \gamma\,TV(x_p) \\ &+ \tfrac{1}{2}\|x - x_p - x_t\|_2^2,\end{aligned} \tag{9}$$

where TV is the Total Variation. To solve this optimization problem, we use the iterative shrinkage algorithm introduced by J. Starck et al. [31]. Once the separate estimates $\hat{x}_p$ and $\hat{x}_t$ are obtained as a solution of (9), the final estimator of x is $\hat{x} = \hat{x}_p + \hat{x}_t$. With multi-channel imagery such as HSI, we can carry out this separation independently per channel (per individual frame corresponding to each spectral wavelength).

[3]This last formulation is known in statistics as penalized least square estimation problem.

C. Multi-Task Joint Sparse Representation

Variants of the sparse representation based classifier have been developed for use with data with multiple dictionaries [33] — this work has been motivated by applications such as multi-modality data fusion, where different views generate different dictionaries that represent the same underlying classification task. We will employ a recent variant of this approach, the multi-task joint sparse representation classifier as our backend classifier to demonstrate the efficacy of morphologically decoupled multi-scale sparse representation. In this approach, multi-source data are jointly represented by a sparse linear combination of the training data across the multiple dictionaries in the ensemble. To learn the joint sparsity of coefficients, the goal is to obtain a row-sparse coefficient matrix which can be modeled as an ℓ_1/ℓ_q-regularized least square problem. For a test sample y^j from source j, given the dictionary $\{\mathcal{A}^j\}_{j=1}^M$ for M sources, the joint sparse coefficient $\mathcal{S} = [\beta^1, \beta^2, ..., \beta^M] \in \mathbb{R}^{n \times M}$ can be estimated by

$$\hat{\mathcal{S}} = \arg\min_{\mathcal{S}} \sum_{j=1}^{M} \left\| y^j - \mathcal{A}^j \beta^j \right\|_2^2 + \lambda \|\mathcal{S}\|_{1,q}, \tag{10}$$

where $\|\mathcal{S}\|_{1,q}$ is the ℓ_1/ℓ_q norm defined as $\|\mathcal{S}\|_{1,q} = \sum_{k=1}^{n} \left\| r^k \right\|_q$, where r^k are the row vectors of $\mathcal{S}$. To make the function convex, q is often set to be greater than 1 (typically 2). Solving the resulting ℓ_1/ℓ_q optimization problem results in a sparse coefficient matrix has common support at the column level.

The problem in (10) is convex but non-smooth. An alternating direction method of multipliers (ADMM) [34], [35] is used to solve this optimization problem. The problem in (10) can be reformulated via the variable splitting technique, i.e., splitting $\mathcal{S}$ into two variables by introducing an auxiliary variable $\mathcal{V}$ as

$$\arg\min_{\mathcal{S},\mathcal{V}} L(\mathcal{S}) + \lambda \|\mathcal{V}\|_{1,2} \quad \text{s.t. } \mathcal{S} = \mathcal{V}, \tag{11}$$

where $L(\mathcal{S}) = \sum_{j=1}^{M} \left\| y^j - \mathcal{A}^j \beta^j \right\|_2^2$. The resulting constrained optimization problem leads to a more tractable solution to the non-smooth problem.

To solve this equality constrained problem, the augmented Lagrangian method can be applied as follows

$$\arg\min_{\mathcal{S},\mathcal{V}} L(\mathcal{S}) + \lambda \|\mathcal{V}\|_{1,2} + \frac{\nu}{2} \|\mathcal{S} - \mathcal{V} + \frac{1}{\nu}\mathcal{B}\|_F^2, \tag{12}$$

where $\mathcal{B}$ is the Lagrangian multiplier, and ν is a positive penalty parameter which is used as the step size during updates.

The problem in (12) can be solved in an alternating fashion [36] — update one variable while fixing the others. In other words, it updates $\mathcal{S}$ while fixing $\mathcal{V}$ and $\mathcal{B}$. In the next iteration, $\mathcal{V}$ is updated while $\mathcal{S}$ and $\mathcal{B}$ are fixed, and so on for $\mathcal{B}$. This process is terminated when an appropriate stopping criterion

is met (in our implementation, when the change in objective function with successive iteration becomes smaller than a preset threshold).

Once $\hat{\mathbb{S}}$ is obtained, the class label associated with a test sample is decided by the total minimal residual

$$\omega = \underset{l=1,2,\ldots,c}{\arg\min} \sum_{j=1}^{M} \left\| y^j - \mathcal{A}^j \delta_l(\hat{\beta}^j) \right\|_2^2 \tag{13}$$

where δ_l denotes an indicator function for the l^{th} class — it ensures that only coefficients $\hat{\beta}^j$ that correspond to atoms from the l^{th} class contribute to the residual. Henceforth, we assume that we have c classes in our dictionary and the image. We remark that this approach is particularly suitable to the proposed morphologically decoupled multi-scale framework wherein the image is partitioned into key texture and cartoon components, resulting in M sub-dictionaries $\{\mathcal{A}^j\}_{j=1}^{M}$ for the hyperspectral image being analyzed. We note that in principle, the proposed framework can utilize (and will be effective for) any sparse representation based classifier at the backend, not just this approach that we chose to validate our framework in this paper.

III. Proposed approach

A. *Morphologically Decoupled Multi-Scale Sparse Representations (MDSR)*

In this section, we describe our proposed MDSR approach in detail. Let j and ℓ denote the scale and direction in the shearlet transform, and j^a and j^f further denote the coarse and fine scales respectively. m denotes the image dimensionality (number of spectral channels) and N_1 represents the number of available training samples. The proposed algorithm is described in Algorithm 1. In the first step, the MCA operation is undertaken independently on each spectral channel of the hyperspectral image. This provides two types of dictionaries for SRC based classification: Dictionaries corresponding to shearlet coefficients at different scales and orientations, representing the cartoon like properties of the image, and dictionaries corresponding to texture features (derived from the recovered DCT image). We would like to point out that our use of analysis coefficients for shearlet coefficients, and synthesized texture images is deliberate. By working with shearlet coefficients directly for classification, we can potentially obtain orientation invariance in classification (in addition to noise robustness) with an appropriate design of the classifier. On the contrary, with regards to the texture part of an image, the synthesized texture image contains image specific texture descriptors as opposed to the raw DCT coefficients which do not carry any information that spatially correlates with information in the original image, thereby being unsuitable in the proposed approach.

From MCA, one builds an ensemble of dictionaries — $\mathcal{A}_t$ representing texture components, and $\{\mathcal{A}_p\}^{j\ell}$ representing cartoon like components via shearlet coefficients at scale s and orientation d respectively. This sets up our multi-task joint-sparse representation model, where a test sample is simultaneously represented in each of these decoupled components individually, resulting in a weighted global residual over these views. The *min* operation, $\tilde{r}^l_{j^f} = \min_d(\, r^l_{j^f\ell})$ that minimizes residuals across all orientations ℓ at each fine scale j_f is crucial to imparting orientation invariance in the proposed framework. The overall class membership function computes a weighted sum of residuals across the texture and approximation dictionaries, and the minimal residual across orientations at each scale for the fine-scale components. The weighting factors for each dictionary $\{w_{j^a}, w_{j^f\ell}, w_t\}$ are estimated as a Fisher's like ratio of between class to within class reconstruction errors

$$
\begin{aligned}
E_j^{(\mathrm{w})} &= \frac{1}{N_1}\sum_{l=1}^{c}\sum_{i\in\text{class-}l} \|a_i - \mathcal{A}_j\delta_l(\hat{\beta})\|^2, \\
E_j^{(\mathrm{b})} &= \frac{1}{N_1(c-1)}\sum_{l=1}^{c}\sum_{i\in\text{class-}l}\sum_{z\neq l} \|a_i - \mathcal{A}_j\delta_z(\hat{\beta})\|^2, \\
w_j &= \frac{E_j^{(\mathrm{b})}}{E_j^{(\mathrm{w})}}.
\end{aligned} \tag{14}
$$

where a_i is i-th atom in the dictionary $\mathcal{A}$ and c is the number of classes. These weights scale the residual associated with SRC from each dictionary such that dictionaries that are more discriminative are given preference in the overall decision function.

B. Rotational Invariance

Next, we show that shearlet coefficients are shear-covariant and approximately rotation-covariant. This property is useful to construct features that are approximately rotation-invariant, and plays a key role in the proposed approach where the minimum residual over all orientations is retained at the fine shearlet scales.

Recall that a representation T of a function f is *covariant* to the action of a group G if the action of any $g \in G$ produces a corresponding shift in the coefficients, that is,

$$T(g \cdot f) = g \cdot T(f).$$

Let D_M denote the dilation operator with respect to an invertible matrix M, that is

$$(D_M)f(x) = |\det M|^{1/2} f(Mx).$$

A direct calculation (using the notation from Sec. II-A) gives the following equalities

$$\begin{aligned}\langle (D_{B_\nu}^{\ell'})f, \psi_{j,\ell,k}^{\nu}\rangle &= \langle f, (D_{B_\nu}^{-\ell'})\psi_{j,\ell,k}^{\nu}\rangle \\ &= \langle f, (D_{B_\nu}^{-\ell'} D_{A_\nu}^{j} D_{B_\nu}^{\ell})\psi^{\nu}\rangle \\ &= \langle f, (D_{A_\nu}^{j} D_{B_\nu}^{\ell-2^j\ell'})\psi^{\nu}\rangle \\ &= \langle f, \psi_{j,\ell-2^j\ell',k}^{\nu}\rangle,\end{aligned}$$

showing that the shearing of f produces a shift of the shearing parameter ℓ of the shearlet coefficients. Hence shearlet coefficients are shear-covariant.

For relatively small angles, a rotation is well approximated by a shearing transformation. In fact, as shown in Sec. II-A, the shearlet transform is defined by restricting the shearing transformation over two cones in such a way to produces directional filters arranged over a pseudo-polar grid, a common approach to approximate a polar grid in the discrete setting [37], [38]. It follows that the shear-covariant shearlet coefficients are also approximately rotation-covariant.

The shear-covariance of the shearlet coeffients can be used to defined a shear-invariant feature of f such as the quantities

$$\max_{\ell}\langle f, \psi_{j,\ell,k}^{\nu}\rangle \text{ or } \min_{\ell}\langle f, \psi_{j,\ell,k}^{\nu}\rangle.$$

By the observation above, these features are also approximately rotation-invariant. This insight leads us to design the classifier such that at each fine scale, the minimum residual over all orientations is retained (c.f. line 7 of Algorithm 1).

To provide insights on the robustness of MDSR to rotational variations between training and testing data with real hyperspectral image data, we show results in Fig. 2 with a commercial building class (that has many different orientations throughout the scene) from a benchmarking hyperspectral dataset — the University of Houston hyperspectral image (described in sec. IV). Throughout the wide-geographic scene, the building class appear in many different orientations. Fig. 2 depicts class-specific residual (c.f. line 7 of Algorithm 1) for the building class. The figure depicts the true-color image of the building, class-specific residuals for the building class using individual dictionaries comprised of each of the 6 shearlet directions, i.e., $r_{j^f\ell}^{l}$, and class specific residual for the building class using $\tilde{r}_{j^f}^{l} = \min_{\ell}(\ r_{j^f\ell}^{l}\ )$. We note that the approach we propose (minimizing residuals across all orientations) accurately characterizes the structure of the building even though the training data was only comprised of 10 pixels from each class gathered (by randomly sampling) from other locations in the scene (with different orientations for this class).

C. Effect of noise

The sparsity of a signal representation entails the ability to capture the fundamental geometric content of the data and, as a result, to more efficiently remove noise [39]. This principle is widely applied in many successful denoising schemes, such as the celebrated shrinkage-based denoising developed by Donoho and Johnstone [40]–[42], which exploits sparsity by observing that when signals are well approximately using a relatively small number representation coefficients then most of the noise is effectively removed by thresholding the representation coefficients.

Let f_n be a noisy image, that is, $f_n = f + n$, where n is a mean-zero Gaussian white noise of variance σ^2. When using the shearlet transform to represent f_n as

$$f_n = \sum_{\mu} \langle f_n, \tilde{\psi}_\mu \rangle \, \psi_\mu,$$

each shearlet coefficient $c_\mu = \langle f, \tilde{\psi}_\mu \rangle$ is affected by mean-zero Gaussian white noise of a variance

$$\mathbb{E}[|c_\mu|^2] - |\mathbb{E}[c_\mu]|^2 = \sigma^2 \|\tilde{\psi}_\mu\|^2$$

and a covariance between shearlet coefficients c_μ and $c_{\mu'}$ of

$$\mathbb{E}[c_\mu \overline{c_{\mu'}}] = \sigma^2 \langle \tilde{\psi}_\mu, \tilde{\psi}_{\mu'} \rangle.$$

Since the shearlets $\tilde{\psi}_\mu$ are normalized, the variance of each shearlet coefficient is a constant, independent of the scale. This is similar to what already observed for wavelets [43]). However, due to shearlets' ability to provide (nearly) optimally sparse approximations for images with edges, the larger-magnitudes shearlet coefficients (those associated with edges and other data structures) are relatively less affected by the noise. This property leads to features with better signal-to-noise ratio than those derived from more conventional function representations, e.g., wavelets. We refer to [44] for a more detailed theoretical analysis of the properties of sparse overcomplete dictionaries with noisy data.

IV. Experimental Setup and Results

We validate the proposed methods and compare their efficacy with traditional hyperspectral classification approaches using two real world hyperspectral datasets. The first image is acquired using an aerial ITRES-CASI (Compact Airborne Spectrographic Imager) 1500 hyperspectral imager over the University of Houston campus and the neighboring urban area. This geospatial image has spatial dimensions of 1001×281 pixels with a spatial resolution of $2.5m$ per pixel. There are 13 classes and 144 spectral bands over the $380 - 1050nm$ wavelength range, representing common urban classes. Parking lot-1 and

Parking lot-2 represent parking lots with and without cars respectively. This dataset was released by us to the research community via the IEEE data fusion contest[1] and covers a wide geographic area over the city of Houston — as a result, it is a challenging dataset with spectral and spatial variability of the various material classes in the scene.

The second dataset represents a unique "forward looking hyperspectral image" of a natural scene, where we seek to classify typical material classes in a natural scene acquired at the University of Houston campus. The image represents a natural scene with different material types, and is acquired from a Headwall photonics Micro Hyperspec VNIR camera. The image size is 1004×1601 pixels, with 163 spectral bands that densely sample the spectrum over the visible and near infrared spectral range $400nm$ through $1001nm$. With recent technological advances, portable hyperspectral imagers that would be able to be deployed for natural image analysis would become prevalent, and a first step to better image understanding with such imagers would entail robust material classification. With that in mind, we extract 9 material classes from this dataset and setup a hyperspectral classification problem over these 9 classes. The scene contains buildings, vehicles, bicycles, roads, vegetation and sky. In a more general setup, such a library could involve common material types that are specific to the scene understanding task at hand. Material with such images can be particularly useful for emerging for image understanding tasks that involve hyperspectral and multispectral images.

Fig. 3 depicts the datasets along with the mean spectral reflectance profiles of the key material types/classes identified in each dataset. With both datasets, we varied the number of training pixels per class (to study the sensitivity to sample size) as indicated in the various results, while the number of test pixels was fixed to 100 pixels. We randomly sample training and test pixels 10 times from the labeled pool, and report average accuracies over these 10 trials.

We next summarize key algorithmic parameters used in this paper. We used a two scale shearlet decomposition (each with six orientations) per spectral channel. A Grey Level Co-occurence Matrix (GLCM) based texture feature extractor was utilized for φ in step 2 of Algorithm 1. Specifically, texture features (contrast, entropy, correlation, energy, homogeneity and variance) are extracted over a window around each pixel, with a window size (determined empirically to be the best window size for each dataset) of 11×11 for the University of Houston aerial data, and 19×19 for the forward looking ground data respectively (the forward looking hyperspectral image has much finer spatial resolution). The values of λ and ν in (12) are set to 0.01, and γ in (9) is set to 500, determined empirically via cross-validation.

[1] http://hyperspectral.ee.uh.edu/?page_id=459

A. Class Dependent Texture

In this experiment, we demonstrate the benefit of morphologically decoupled ensemble of dictionaries, by comparing classification performance of the proposed system with and without the texture component. Specifically, we provide class-specific accuracies of MDSR with and without the texture specific dictionary in the ensemble in Table I. We note that classes with a significant texture contribution (e.g. trees, residential buildings, etc.) see a substantial benefit when the texture component is included in the ensemble of dictionaries. This further underscores the premise of this work — that decomposing an image into texture and cartoon components and building appropriate dedicated dictionaries for each component results in a very robust representation with regards to image classification.

TABLE I: Average classification accuracy for individual classes in the UH dataset, with and without texture

Sample Size	**Training Sample Size 5**		**Training Sample Size 10**	
Class Name / Algorithm	**MDSR**	**MDSR (No texture)**	**MDSR**	**MDSR (No texture)**
Grass-healthy	77.1	77.5	87.7	83.9
Grass-stressed	84.1	84.3	87.2	91.3
Grass-synthetic	95.6	89.1	96.5	92.4
Tree	86.7	63.4	94.4	79.3
Soil	98.2	98.3	99.3	99.7
Water	92.4	90.9	93.4	91.4
Residential	81.6	80.1	89.7	83.8
Commercial	51.8	59.8	71.7	76.6
Road	77.3	54.6	89.9	67.8
Parking Lot 1	76.5	65.0	83.2	75.5
Parking Lot 2	97.4	96.7	99.3	99.7
Tennis Court	97.9	99.8	100.0	100.0
Running Track	94.0	93.8	97.4	95.3
Overall Accuracy	85.4	81.0	91.5	87.4

B. Noise Robustness

In this experiment, we validate and quantify the noise robustness of the proposed morphologically decoupled image analysis approach. Specifically, we simulate noisy hyperspectral imagery at different signal to noise ratios by adding white Gaussian noise. We then compare the performance of the proposed

classification approach (which does not require any explicit denoising), with traditional SRC classifiers wherein the data has been denoised with a Wiener filter (along the spectral and spatial dimensions of the hyperspectral cube respectively). The Wiener filtered accuracies are obtained by employing a Wiener filter as a preprocessing to the image along the spatial dimensions (independently per spectral wavelength) and along the spectral dimension (per pixel) — A single SRC classifier is then trained and validated on the hyperspectral data, with spectral reflectance based features. Results from this experiment are depicted in Fig. 4. As expected, a spatial Wiener filter outperforms the spectral Wiener filter, while the proposed approach offers a much more noise robust performance. We note the substantially high classification accuracy with the proposed approach, compared to a single sparse representation classifier that operates on Wiener filter based denoised imagery, in addition to the much slower drop in performance as PSNR decreases.

C. Rotational Invariance and Sensitivity to Sample Size

TABLE II: Classification Accuracy with the University of Houston (airborne) Hyperspectral Image (average overall accuracies along with standard deviations in parenthesis).

Algorithm / Sample Size		**5**	**10**	**15**	**20**
Proposed (MDSR)	**MD, MS, RI** & **Weighted**	85.43 (1.63)	91.52 (1.44)	93.44 (1.63)	95.12 (1.20)
	MD, MS & **RI**	84.55 (1.72)	90.91 (1.73)	93.28 (1.62)	95.08 (1.10)
	MD, MS	77.82 (1.99)	87.68 (2.11)	91.50 (1.50)	93.72 (1.49)
Baseline	**raw-spectral**	81.52 (2.87)	86.31 (1.76)	86.89 (1.14)	86.98 (1.25)

TABLE III: Classification Accuracy with the forward looking Hyperspectral Image of a natural scene (average overall accuracies along with standard deviations in parenthesis).

Algorithm / Sample Size		**5**	**10**	**15**	**20**
Proposed (MDSR)	**MD, MS, RI,** & **Weighted**	85.81 (2.48)	94.19 (2.06)	96.69 (0.99)	97.30 (0.69)
	MD, MS & **RI**	78.80 (2.61)	89.82 (2.04)	94.80 (1.16)	96.58 (0.82)
	MD, MS	65.97 (1.05)	79.61 (2.10)	85.44 (1.81)	88.30 (1.53)
Baseline	**raw-spectral**	82.18 (2.84)	86.80 (1.83)	88.14 (1.38)	87.86 (1.11)

In this experiment, we demonstrate the rotational invariance property of the proposed method, and illustrate performance as a function of training sample size. Specifically, we show results by adding the various components in the proposed framework sequentially (morphological decoupling, multi-scale analysis, rotational invariance and adaptive scaling of residuals). We also provide comparison to a single SRC classifier that is built on the raw spectral reflectance features. With the proposed framework, we present results with three variations: Morphologically Decoupled, Multi-Scale (*MD, MS*), Morphologically Decoupled, Multi-Scale and Rotational Invariant (*MD, MS* & *RI*), and Morphologically Decoupled, Multi-Scale, Rotational Invariant and Weighted residuals (*MD, MS, RI,* & *Weighted*). *MD, MS* connotes a multi-task SRC implementation wherein each scale and orientation of the shearlet coefficients, along with texture features form a dedicated dictionary, and the final classification decision is made by minimizing the sum of residuals over all these dictionaries. In the *MD, MS* & *RI* approach, instead of accumulating residuals across all orientations and scales, for each scale, we pick the smallest residual over all possible orientations. These "minimum residuals over all orientations" across the various shearlet scales (and texture) are then summed up. In the final variant of the proposed method, *MD, MS, RI,* & *Weighted*, we weigh individual dictionaries by weights that reflect their relative discriminative ability for the classification task. Results for both datasets are summarized in Table II and Table III. For both datasets, the proposed MDSR approach substantially outperforms classification using spectral reflectance features only. The weighted variant of the proposed approach (MDSR: *MD, MS, RI,* & *Weighted*), results in the best overall performance. Picking the minimum residual over all orientations has a profound impact in the underlying multi-task sparse representation task. This is due to the rotational invariance brought about by this minimization — we assert that this helps account for objects that possess different relative orientations in the test samples and the training dictionaries. An appropriate weighting of these residuals further boosts classification performance.

V. Conclusions and Future Work

In this paper, we presented an approach to morphologically decoupled multi-scale image classification. The approach is demonstrated to be very effective for sparse representation based classifiers, and is particularly suited to hyperspectral data — the high dimensionality of hyperspectral data, coupled with limited training data make such classifiers particularly suited to such data. The morphological decoupling and appropriate treatment of orientation-specific dictionaries results in an image classification framework that is robust to additive noise, limited training sample size, rotational variabilities between training and test samples. We provided theoretical and intuitive insights into the benefits of such a framework.

We also validated the framework with two datasets — an aerial hyperspectral image, and a forward looking hyperspectral image representing a *natural* scene, and results with these datasets demonstrates the competitive and robust classification performance of the proposed framework. This approach is very well suited to classify data that can be modeled as a linear superposition of texture and cartoon-like components — an assumption that fits most natural and geospatial images.

REFERENCES

[1] S. Prasad, P. Gamba, and M. Herold, "Foreword to the special issue on earth observation approaches for large area land monitoring with multiple sensors and resolutions," *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, vol. 6, no. 5, pp. 2075–2076, Oct 2013.

[2] D. Landgrebe, "Hyperspectral image data analysis," *Signal Processing Magazine, IEEE*, vol. 19, no. 1, pp. 17–28, 2002.

[3] D. Manolakis and G. Shaw, "Detection algorithms for hyperspectral imaging applications," *Signal Processing Magazine, IEEE*, vol. 19, no. 1, pp. 29–43, 2002.

[4] G. Shaw and D. Manolakis, "Signal processing for hyperspectral image exploitation," *Signal Processing Magazine, IEEE*, vol. 19, no. 1, pp. 12–16, 2002.

[5] Z. Pan, G. Healey, M. Prasad, and B. Tromberg, "Face recognition in hyperspectral images," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 12, pp. 1552–1560, 2003.

[6] G. Camps-Valls and L. Bruzzone, "Kernel-based methods for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 6, pp. 1351–1362, June 2004.

[7] N. Keshava and J. Mustard, "Spectral unmixing," *Signal Processing Magazine, IEEE*, vol. 19, no. 1, pp. 44–57, 2002.

[8] C. Li, T. Sun, K. F. Kelly, and Y. Zhang, "A compressive sensing and unmixing scheme for hyperspectral data processing," *Image Processing, IEEE Transactions on*, vol. 21, no. 3, pp. 1200–1210, 2012.

[9] D. W. J. Stein, S. G. Beaven, L. E. Hoff, E. M. Winter, A. P. Schaum, and A. D. Stocker, "Anomaly detection from hyperspectral imagery," *IEEE Signal Processing Magazine*, vol. 19, no. 1, pp. 58–69, January 2002.

[10] D. W. J. Stein, S. Beaven, L. Hoff, E. Winter, A. Schaum, and A. Stocker, "Anomaly detection from hyperspectral imagery," *Signal Processing Magazine, IEEE*, vol. 19, no. 1, pp. 58–69, 2002.

[11] S. Prasad and L. M. Bruce, "Limitations of principal component analysis for hyperspectral target recognition," *IEEE Geoscience and Remote Sensing Letters*, vol. 5, no. 4, pp. 625–629, October 2008.

[12] M. Cui and S. Prasad, "Angular discriminant analysis for hyperspectral image classification," *Selected Topics in Signal Processing, IEEE Journal of*, vol. PP, no. 99, pp. 1–1, 2015 (accepted).

[13] D. Lunga, S. Prasad, M. Crawford, and O. Ersoy, "Manifold-learning-based feature extraction for classification of hyperspectral data: A review of advances in manifold learning," *Signal Processing Magazine, IEEE*, vol. 31, no. 1, pp. 55–66, 2014.

[14] Y. Chen, N. Nasrabadi, and T. Tran, "Sparse representation for target detection in hyperspectral imagery," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 5, no. 3, pp. 629–640, 2011.

[15] C. Lan, X.-Y. Jing, S. Li, L. Bian, and Y.-F. Yao, "Exploring the natural discriminative information of sparse representation for feature extraction," in *Image and Signal Processing (CISP), 2010 3rd International Congress on*, vol. 2, 2010, pp. 916–920.

[16] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 471–478.

[17] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 2, pp. 210–227, 2009.

[18] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, and Y. Ma, "Towards a practical face recognition system: Robust registration and illumination by sparse representation," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2009, pp. 597–604.

[19] J. Yang, J. Wright, T. Huang, and Y. Ma, "Image super-resolution via sparse representation," *Image Processing, IEEE Transactions on*, vol. 19, no. 11, pp. 2861–2873, 2010.

[20] Z. Zhou, A. Wagner, H. Mobahi, J. Wright, and Y. Ma, "Face recognition with contiguous occlusion using markov random fields," in *Computer Vision, 2009 IEEE 12th International Conference on*, 2009, pp. 1050–1057.

[21] K. Guo and D. Labate, "Geometric separation of singularities using combined multiscale dictionaries," *J. Fourier Anal. Appl. (in press)*, 2015.

[22] D. L. Donoho and G. Kutyniok, "Microlocal analysis of the geometric separation problem," *Commun. Pure Appl. Math.*, vol. 66, no. 1, pp. 1–47, 2013. [Online]. Available: http://dx.doi.org/10.1002/cpa.21418

[23] M. Elad, J. L. Starck, P. Querre, and D. L. Donoho, "Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA)," *Applied and Computational Harmonic Analysis*, vol. 19, no. 3, pp. 340–358, November 2005. [Online]. Available: http://dx.doi.org/10.1016/j.acha.2005.03.005

[24] G. R. Easley, D. Labate, and P. Negi, "3D data denoising using combined sparse dictionaries." *Math. Model. Nat. Phenom.*, vol. 8, no. 1, pp. 60–74, 2013.

[25] J.-L. Starck, M. Elad, and D. L. Donoho, "Redundant multiscale transforms and their application for morphological component analysis," *Adv. Imag. Electron Phys.*, vol. 132, pp. 287–348, 2004.

[26] G. Kutyniok and D. Labate, *Shearlets: Multiscale Analysis for Multivariate Data*. Springer, 2012.

[27] D. Labate, W. Lim, G. Kutyniok, and G. Weiss, "Sparse multidimensional representation using shearlets," *SPIE Proc. 5914, SPIE, Bellingham*, pp. 254–262, 2005. [Online]. Available: http://en.wikipedia.org/wiki/Bandelet_(computer_science)

[28] E. J. Candès and D. L. Donoho, "New tight frames of curvelets and optimal representations of objects with piecewise C^2 singularities." *Commun. Pure Appl. Anal.*, vol. 57, no. 2, pp. 219–266, 2004.

[29] K. Guo and D. Labate, "The construction of smooth Parseval frames of shearlets." *Math. Model. Nat. Phenom.*, vol. 8, no. 1, pp. 82–105, 2013.

[30] ——, "Optimally sparse multidimensional representation using shearlets," *SIAM J. Math. Analysis*, vol. 39, no. 1, pp. 298–318, 2007.

[31] J. Starck, M. Elad, and D. L. Donoho, "Image decomposition via the combination of sparse representations and a variational approach," *IEEE Transactions on Image Processing*, vol. 14, no. 10, pp. 1570–1582, 2005.

[32] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1998.

[33] X.-T. Yuan, X. Liu, and S. Yan, "Visual classification with multitask joint sparse representation," *Image Processing, IEEE Transactions on*, vol. 21, no. 10, pp. 4349–4360, 2012.

[34] J. Yang and Y. Zhang, "Alternating direction algorithms for \ell_1-problems in compressive sensing," *SIAM journal on scientific computing*, vol. 33, no. 1, pp. 250–278, 2011.

[35] M. V. Afonso, J. M. Bioucas-Dias, and M. A. T. Figueiredo, "An augmented Lagrangian approach to the constrained optimization formulation of imaging inverse problems," *IEEE Transactions on Information Theory*, vol. 20, no. 3, pp. 681–695, March 2011.

[36] S. Shekhar, V. M. Patel, N. M. Nasrabadi, and R. Chellappa, "Joint sparse representation for robust multimodal biometrics recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 1, pp. 113–126, 2014.

[37] A. Averbuch, R. R. Coifman, D. L. Donoho, M. Israeli, and Y. Shkolnisky, "A framework for discrete integral transformations i—the pseudopolar fourier transform," *SIAM Journal on Scientific Computing*, vol. 30, no. 2, pp. 764–784, 2008.

[38] G. R. Easley, D. Labate, and W. Lim, "Sparse directional image representations using the discrete shearlet transform," *Appl. Comput. Harmon. Anal.*, vol. 25, pp. 25–46, 2008.

[39] D. L. Donoho, M. Vetterli, R. A. DeVore, and I. Daubechies, "Data compression and harmonic analysis," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2435–2476, 1998.

[40] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 613–627, 1995.

[41] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994.

[42] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard, "Wavelet shrinkage: asymptopia," *Journal of the Royal Statistical Society, Ser. B*, pp. 371–394, 1995.

[43] S. Mallat, *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*, 3rd ed. Academic Press, 2008.

[44] M. Raphan and E. P. Simoncelli, "Optimal denoising in redundant representations," *IEEE Transactions on Image Processing*, vol. 17, no. 8, pp. 1342–1352, 2008. [Online]. Available: http://dx.doi.org/10.1109/TIP.2008.925392

Algorithm 1 *MDSR*

Input: A vectorized m-dimensional image $x \in \mathbb{R}^{N^2 \times m}$, test pixel $y \in \mathbb{R}^m$.

{***Morphological Separation***}

for all $i \in 1, 2, \ldots, m$ **do**

- Calculate the shearlet and DCT coefficients for x^i ($i-$th column of x) based on MCA:

$$\{\ \{\hat{\alpha}_p^i\}^{j\ell},\ \hat{\alpha}_t^i\ \} = \mathit{MCA}\ (\ x^i,\ \mathcal{D}_p,\ \mathcal{D}_t\ ).$$

- Generate the shearlet coefficient matrix for each scale j and each direction ℓ: $\{\mathcal{C}_p^i\}^{j\ell} = \{\hat{\alpha}_p^i\}^{j\ell}$.
- Recover the DCT texture image: $\hat{x}_t^i = \mathcal{D}_t \hat{\alpha}_t^i$.
- Extract texture features from $\hat{x}_t^i$: $\tilde{x}_t^i = \varphi(\hat{x}_t^i)$, where φ denotes a textural feature extractor.

end for

{***Sparse Representation over Ensemble of Dictionaries***}

Assume $\{\mathcal{A}_p \in \mathbb{R}^{N_1 \times m}\}^{j\ell}$ and $\mathcal{A}_t \in \mathbb{R}^{N_1 \times m}$ are the training dictionaries generated from $\{\mathcal{C}_p\}^{j\ell}$ and $\tilde{x}_t$.

Obtain representation coefficients ($\{\ \{\hat{\beta}_p\}^{j\ell},\ \hat{\beta}_t\ \}$) corresponding to each dictionary based on (10) and (12).

{***Morphologically Decoupled Classification***}

Compute residuals: For the test pixel y for l-th class:

$$r_{j\ell}^l = \|y - \{\mathcal{A}_p\}^{j\ell}\delta_l(\{\hat{\beta}_p\}^{j\ell})\|_2,$$
$$r_t^l = \|y - \mathcal{A}_t\delta_l(\hat{\beta}_t)\|_2.$$

Rotation invariance: Calculate the minimum residuals of fine scales s^f with regard to different directions d:

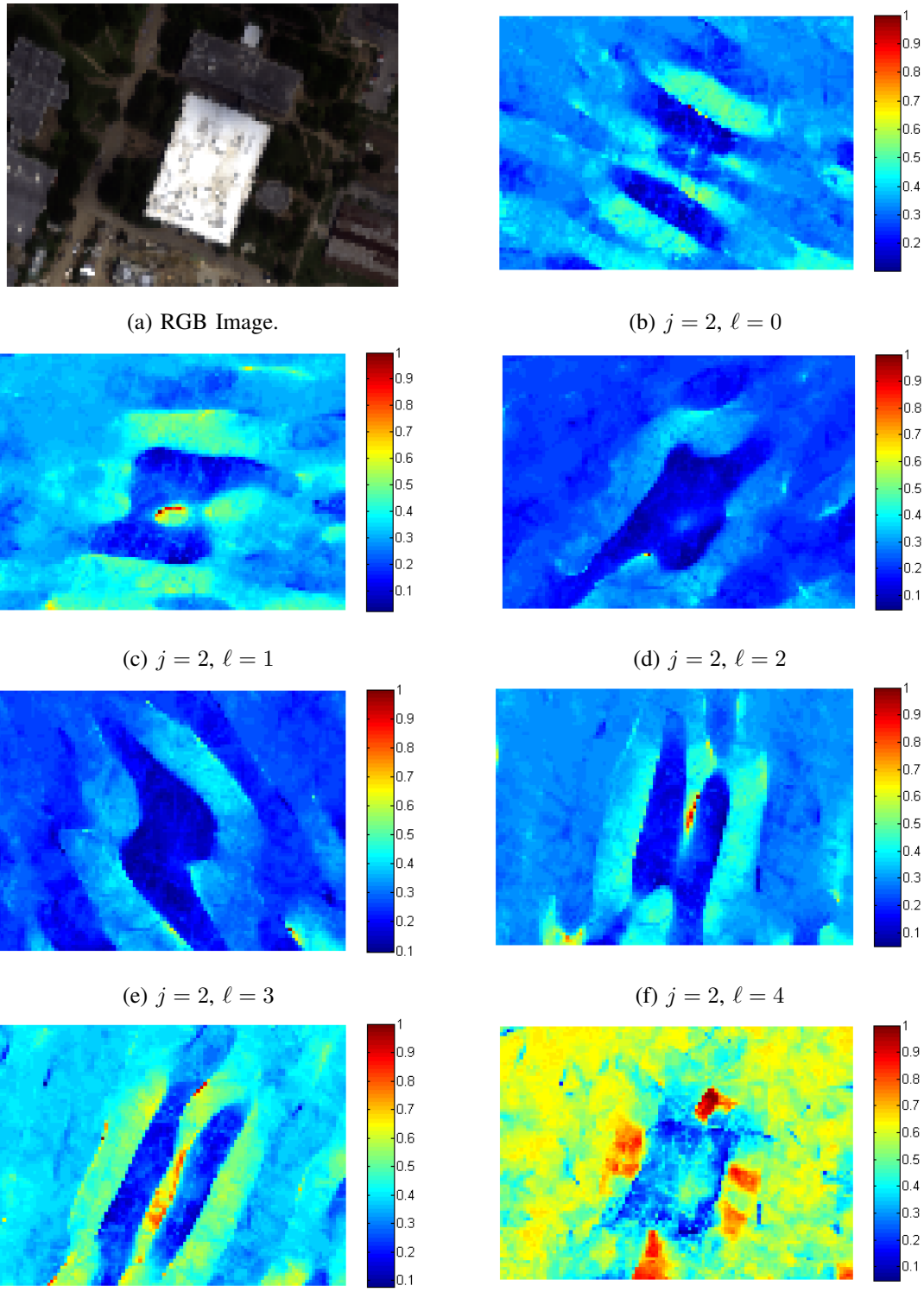
$$\tilde{r}_{j^f}^l = \min_{\ell}(\ r_{j^f\ell}^l\ ).$$

Adaptive weighting of residuals: Use (14) to estimate scaling of residuals corresponding to every dictionary.

Classification: Determine the class label of a test pixel y based on:

$$\omega = \underset{l=1,2,\ldots,c}{\operatorname{argmin}}(w_{j^a} r_{j^a}^l + \sum_{j^f\ell} w_{j^f\ell}\tilde{r}_{j^f\ell}^l + w_t r_t^l).$$

Output: A class label ω.



(a) RGB Image.

(b) $j = 2$, $\ell = 0$

(c) $j = 2$, $\ell = 1$

(d) $j = 2$, $\ell = 2$

(e) $j = 2$, $\ell = 3$

(f) $j = 2$, $\ell = 4$

(g) $j = 2$, $\ell = 5$

(h) MDSR

Fig. 2: **Residual for the building class for a small cropped portion from the UH dataset, cropped over one of the many buildings in the scene (shown as a natural color image in a), using dictionaries comprised of recovered shearlet coefficients ($r^l_{jf\ell}$) across individual directions (b—g), and using the approach used in MDSR, ($\tilde{r}^l_{jf} = \min_\ell(r^l_{jf\ell})$) — finding the minimum residual across all orientations (h).**

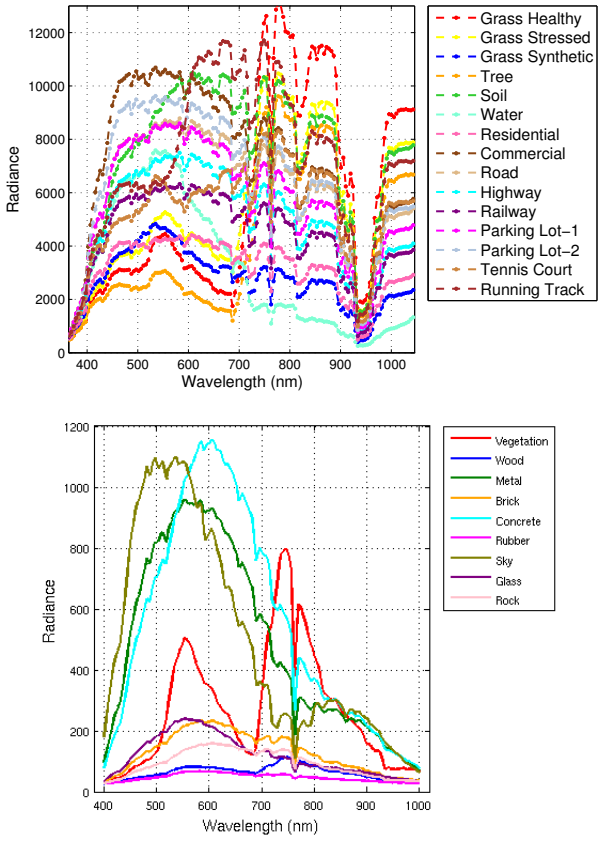


Fig. 3: **Illustrating the spectral reflectance signatures for the University of Houston dataset (left), and the natural scene (right).**

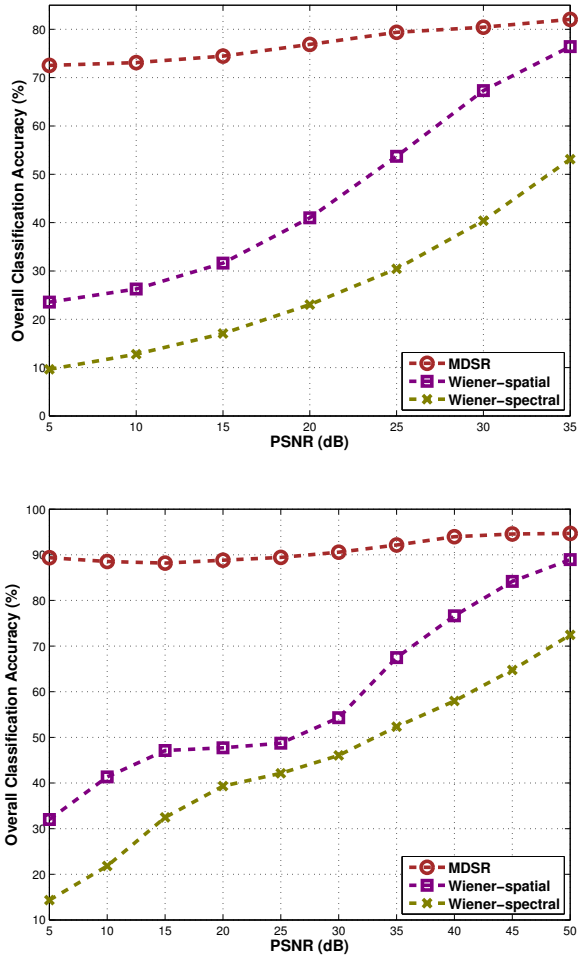


Fig. 4: **Illustrating the overall classification accuracy as a function of PSNR (dB) for the University of Houston dataset (left), and the natural scene (right).**