

# Discriminant Absorption Feature Learning for Material Classification

Zhouyu Fu, Antonio Robles-Kelly, *Senior Member, IEEE*

## Abstract

In this paper, we develop a novel approach to object material identification in spectral imaging by combining the use of invariant spectral absorption features and statistical machine learning techniques. Our method hinges in the relevance of spectral absorption features for material identification and casts the problem into a pattern recognition setting by making use of an invariant representation of the most discriminant band-segments in the spectra. Thus, here, we view the identification problem as a classification task, which is effected based upon those invariant absorption segments in the spectra which are most discriminative between the materials under study. To robustly recover those bands that are most relevant to the identification process, we make use of discriminant learning. To illustrate the utility of our method for purposes of material identification, we perform experiments on both terrestrial and remotely sensed hyperspectral imaging data and compare our results to those yielded by an alternative.

## Index Terms

Hyperspectral Image Analysis, Classification, Absorption Band Detection, Photometric Invariance, Feature Selection/Extraction

## I. INTRODUCTION

Hyperspectral sensing technology can capture image data in tens or hundreds of bands covering a broad spectral range. Compared to traditional monochrome and trichromatic cameras, hyperspectral image sensors can provide an information-rich representation of the spectral response of materials which poses great opportunities and challenges on material identification [1], which is, in general, a classification problem.

For hyperspectral image classification, each pixel is associated with a spectrum which can be viewed as a input vector in a high dimensional space. Thus algorithms from statistical pattern recognition and machine learning have been adopted to perform pixel-level feature extraction and classification [2]. These methods either directly use the complete spectra, or often make use of preprocessing and dimensionality reduction steps at input and attempt to recover statistically optimal solutions. Linear dimensionality reduction methods are based on the linear projection

Z. Fu is with the Gippsland School of Information Technology, Monash University, Churchill, Victoria 3842, Australia and A. Robles-Kelly is with the Canberra Lab of National ICT Australia, Canberra, ACT 2601, Australia and also affiliated with the Department of Information Engineering, Research School of Information Sciences and Engineering, Australian National University

of the input data to a lower dimensional feature space. Typical methods include Principal Component Analysis (PCA) [3], Linear Discriminant Analysis (LDA) [4], and Projection Pursuit [5]. Almost all linear feature extraction methods can be kernelised, resulting in Kernel PCA [6], Kernel LDA [7] and Kernel Projection Pursuit [8]. These methods exploit nonlinear relations between different segments in the spectra by mapping the input data onto a high dimensional space through different kernel functions [9].

From an alternative viewpoint, absorption features can be used as signatures for chemicals and their concentrations [10]. Absorption and reflection are two complementary behaviours of the light incident on the material surface. While reflections are directly measurable by imaging sensors, absorptions, indicated by local dips or valleys in the reflectance spectra, are less straightforward to recover. Nevertheless, absorptions are inherently related to the material chemistry as well as other physical properties such as surface roughness [11]. Therefore, the presence of an absorption at certain spectral range is a “signature”, which can be used for identification and recognition purposes. Furthermore, absorption features have been used in the Tetracorder system [12] to identify spectrum components. Unfortunately, the Tetracorder is a semi-automatic expert system where, for purposes of material identification, absorption features are required to be manually labelled. The Tetracorder is based upon spectral feature identification algorithms such as that in [13], where a least-squares fit is used to match the spectrum under study to that in a reference library. This is akin to linear mixing models [14], [15], [16], where the pixel spectrum is represented as a linear mixture of end-member spectra.

The use of absorption features opens up the possibility of making use of conventional classifiers such as the K-nearest neighbour (KNN) or maximum likelihood (MLC) classifiers for material identification [2]. Recently, Support Vector Machines (SVMs) [17] have been used for classification of remotely sensed data [18], [19], [20], [21], [22], [23].

Further, using the spectra as input to the SVM classifier achieves better performance than that yielded by using lower-dimensional features [19], [24]. This is not surprising, since feature extraction may potentially lose discriminant information, whereas the direct application of SVMs to the spectra can be viewed as an implicit feature selection functionality effected through the kernel mapping.

Hence, statistical and absorption-based approaches are complementary in nature with respect to their respective strengths and limits. There has been little work, to our knowledge, linking together the statistical analysis of the spectra with the domain knowledge of material physics and chemistry. In this paper, we investigate the use of statistical learning approaches for purposes of material identification based upon absorption features. Moreover, the use of absorption features has made it feasible to train statistical classifiers on a small sample set with a marginal detriment on its generalisation properties. The reason is that absorptions are intrinsic to the object under study. The main advantage of this treatment is that, by using absorptions for purposes of identification and recognition, we can perform a localised analysis of the spectra so as to discriminate amongst different spectral segments.

It is important to note that, especially for terrestrial imaging spectroscopy, the geometry of the object under study plays an important role in the identification process. This is due to photometric phenomena, which result from the

image formation process. Healey and his colleagues [25], [26], [27] have addressed the problem of photometric invariance in hyperspectral imagery for invariant material classification and mapping in aerial imaging as related to photometric artifacts induced by atmospheric effects and changing solar illuminations. In [28], a method is presented for hyperspectral edge detection. The method is robust to photometric effects, such as shadows and specularities. In [29], a photometrically invariant approach was proposed based on the derivative analysis of the spectra. This local analysis of the spectra was shown to be intrinsic to the surface albedo. Yet, the analysis in [29] was derived from the Lambertian reflection model and, hence, is not applicable to specular reflections. Fu and Robles-Kelly [24] have proposed the use of band ratios as an alternative to raw spectral bands as features for classification as a means to shading invariance. In [30], a subspace projection method for specularity-free spectral representation is presented.

Despite effective, these methods are not specifically designed for purposes of classification, whereas our primary interest here is to develop methods based on absorption feature learning for better classification performance. Thus, the main contributions of the work proposed here are threefold. Firstly, we present a classification methodology that combines the merits of both absorption modelling and statistical learning techniques for improved pixel-based classification of hyperspectral imagery. Secondly, we develop a novel feature representation scheme that incorporates photometric invariance in the representation of spectral segments truncated from absorption bands and hence is quite robust to varying lighting conditions and photometric artifacts. Finally, we propose a feature selection algorithm that can automatically select the most discriminant absorption segments for purposes of classification based on statistical modelling of dissimilarities of probability distributions of class pairs making use of  $\alpha$ -entropy.

The outline of this paper is as follows. In Section II, we present the general framework of our approach in Section II. In Sections III and IV we describe the main contributions of the paper. Firstly, we present a feature representation for the extracted absorption segments. The representation, despite being simple, is invariant to photometric effects like shading and specularity. Secondly, we present a method for selecting the most discriminant absorption segments based upon  $\alpha$ -entropy and its application to SVM-based classification making use of an invariant absorption feature representation. In Section V, we elaborate further on the algorithm for our method. Section VI provides an experimental evaluation of the method and conclusions are given in Section VII.

## II. PREREQUISITES AND OVERVIEW

As mentioned earlier, our material identification method is based upon a two stage process. The first of these involves the use of absorption features for purposes of training a Support Vector Machine (SVM) classifier. With the trained classifiers at hand, materials can be identified in the testing phase. In Figure 1, we show a diagrammatic representation of our method for feature extraction and classification.

To train the SVM, we commence by recovering the absorption bands for each training spectral sample. Then the absorption segments, i.e. truncated spectrum consisting of spectral values within the absorption bands, are extracted from the training spectra. The purpose is to select the optimal absorption spectral segments that can best

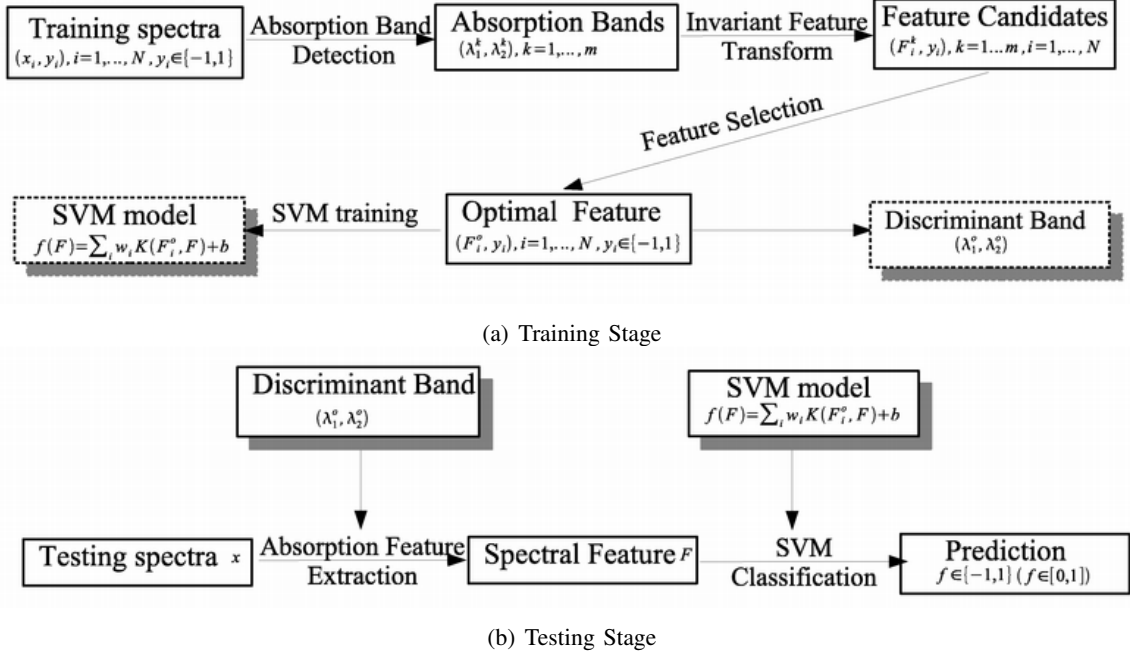


Fig. 1: General Framework for our absorption-based classification algorithm

discriminate between the classes under consideration. As a preprocessing step, we apply a simple transformation to the absorption segments extracted in the detection step. Thus, the candidate features are invariant to various photometric conditions such as shading and specularity. We then proceed to select the absorption features so that the alpha-divergence between the classes under study is maximised.

As a result of the treatment above, we can cast the identification task into an invariant classification setting. To do this, in the testing stage, we truncate the test spectra to the absorption bands of interest and apply our learned feature transformation. We then use the SVM classifier, which has been trained using the scheme described above, for purposes of identification. It is important to stress that, in the testing stage, we do not perform absorption band detection and selection on the testing samples. This is due to the fact the discriminant absorption band is recovered and recorded in the training stage. As a result, at testing, only spectral truncation and ratio operations are involved in obtaining the absorption feature vector, which, in turn, increases computationally efficiency.

### III. ABSORPTION BASED FEATURE REPRESENTATION

As mentioned earlier, we focus our attention in the use of absorptions for material classification in hyperspectral imaging. Along these lines, we commence by noting that, in hyperspectral imaging [31], absorptions can be viewed as local “dips” on the spectra. Mathematically, an absorption can be considered to be a local valley which breaks the convexity of the local spectrum across nearby bands.

As a preliminary step to absorption-based material classification, we need to detect the existing absorption bands in each pixel spectrum. This is so as to identify a set of band edges  $(\lambda_1^j, \lambda_2^j)$  ( $j = 1, \dots, m$ ) with  $\lambda_1^j < \lambda_2^j$ . Each pair of band edges denotes an absorption band. The truncated absorption spectral segment accounts for each absorption

band with spectral values taken from a contiguous range of wavelengths  $\{\lambda | \lambda_1^j \leq \lambda \leq \lambda_2^j\}$  that form a local valley within the full-length spectrum. A number of methods have been proposed to detect and locate absorption bands [32], [33], [34], [35]. Here we have adopted a method proposed by Fu and Robles-Kelly [35], which has been shown to overcome a number of problems elsewhere in the literature [32], [33], [34]. More details on the analysis and empirical evaluation of absorption detection methods can also be found in [35].

With the identified absorption bands at hand, we now present a representation of absorption spectral segments invariant to various photometric conditions. In previous studies [12], [34], absorption features are usually represented in parametric form by a 4-element vector, i.e. the extreme spectral bands, the area and the depth of the absorption under study. However, this representation is an empirical one, which lacks theoretical justification. Moreover, it inevitably loses discriminant information due to the lower order statistics which it comprises.

Hence, in this paper, we represent absorption features directly using the spectral values from the truncated spectral segment whose end-point wavelengths are given by the absorption band edges. From the invariance perspective, we commence by noting that the extracted absorption feature can be rendered invariant to shading through normalisation. To see this, consider the diffuse reflectance model for the hyperspectral image under study. For the pixel location  $u$ , the radiance value at the  $i^{th}$  band with wavelength  $\lambda_i$  is given by

$$I(\lambda, u) = \int_{\lambda} g(u) E(\lambda) S(\lambda, u) Q_i(\lambda) d\lambda \quad (1)$$

where  $E(\lambda)$  is the light source spectral component as a function of the wavelength  $\lambda$ ,  $S(\lambda, u)$  is the surface albedo,  $Q_i(\lambda)$  is the sensor response function for the  $i^{th}$  receptor centered at  $\lambda_i$  and  $g(u)$  is the shading factor at the pixel location  $u$ , which is wavelength-independent and is governed by the illumination and surface geometries.

In spectral imaging, narrow-band receptors are widely applied and, hence, the sensor response for the  $i^{th}$  band can be regarded to be a delta function centered at  $\lambda_i$ . By making use of the shorthand  $Q_i(\lambda) = \delta(\lambda - \lambda_i)$  in Equation 1, we can express the radiance for the  $i^{th}$  band at pixel-site  $u$  as follows

$$I(\lambda_i, u) = g(u) \int_{\lambda} E(\lambda) S(\lambda, u) \delta(\lambda - \lambda_i) d\lambda = g(u) E(\lambda_i) S(\lambda_i, u) \quad (2)$$

To take our analysis further and remove the effects of the illuminant spectra, we normalise the radiance  $I(\lambda_i, u)$  with respect to light source spectrum  $E(\lambda_i)$  at the wavelength of interest. Therefore, for a diffuse reflectance setting, we can write

$$r_u(\lambda_i) = \frac{I(u, \lambda_i)}{E(\lambda_i)} = g(u) S(\lambda_i, u) \quad (3)$$

where  $r_u(\lambda_i)$  is the normalised reflectance spectrum at the pixel-coordinate  $u$ . Due to the presence of the shading factor  $g(u)$ , we still cannot recover the true albedo. Nonetheless, the absorption representation above is independent of the illuminant spectrum  $E(\lambda_i)$ .

Note that, in the equations above, we have not considered specular effects. To account for specularities, we employ the Dichromatic model [36]. This model has long been used in physics-based vision for characterising

specular reflections on non-Lambertian surfaces. Following our assumption of narrow-band receptors, the radiance for a specular pixel can be expressed as follows

$$I(\lambda_i, u) = g(u)E(\lambda_i)S(\lambda_i, u) + k(u)E(\lambda_i) \quad (4)$$

where the second term on the right-hand side of the equation above has been added with respect to Equation 2. This term accounts for the specular reflection. Here,  $k(u)$  is the specular factor which depends on the lighting, surface and viewing geometries at location  $u$ . Note that the specular component is a simple scaling of the illumination spectrum independent of the surface albedo.

Thus, making use of the Dichromatic model above, the analogue of Equation 3 becomes

$$r_u(\lambda_i) = g(u)S(\lambda_i, u) + k(u) \quad (5)$$

Now, we see that the second term in the equation above has been reduced to the specular factor  $k(u)$  which does not depend on the wavelength  $\lambda_i$ . This observation is important since it allows us to eliminate the specular component by taking the difference of reflectance data between two arbitrary bands  $\lambda_i$  and  $\lambda_j$ , i.e.

$$r_u^*(\lambda_i) = r_u(\lambda_i) - r_u(\lambda_j) = g(u)(S(\lambda_i, u) - S(\lambda_j, u)) \quad (6)$$

Theoretically, an arbitrary reference band  $\lambda_j$  can be used for the difference above. Nonetheless,  $r_u(\lambda_i)$  and  $r_u(\lambda_j)$  may deviate from the ideal case as modelled by Equation 5 due to measurement noise. Let  $\hat{r}_u(\lambda_i)$  be the true spectral value at wavelength  $\lambda_i$ , then the measurement  $r_u(\lambda_i)$  is a random variable satisfying the relation  $r_u(\lambda_i) = \hat{r}_u(\lambda_i) + e_{\lambda_i}$ , where  $e_{\lambda_i}$  is the measurement error at wavelength  $\lambda_i$ . Here we assume that the measurement errors  $e_{\lambda_i}$  are independent and identically distributed random variables governed by Gaussian distributions with zero mean and variance  $\sigma^2$ . Denote  $\hat{r}_u^*(\lambda_i) = \hat{r}_u(\lambda_i) - \hat{r}_u(\lambda_j)$  as the true difference value defined in Equation 6. We then have the following results regarding  $r_u^*(\lambda_i)$  and  $\hat{r}_u^*(\lambda_i)$ ,

$$\begin{aligned} E(r_u^*(\lambda_i)) &= E(\hat{r}_u^*(\lambda_i)) \\ \text{var}(r_u^*(\lambda_i) - \hat{r}_u^*(\lambda_i)) &= 2\sigma^2 \end{aligned} \quad (7)$$

Hence the measurement is unbiased with twice as much variance as that in the measurement of individual bands.

With slight modification to Equation 6, we can still eliminate the specular component in Equation 5 while much reducing the variance. This is achieved by using the average of  $r_u(\lambda)$  values over the spectrum instead of an arbitrary reference band

$$r_u^*(\lambda_i) = r_u(\lambda_i) - \frac{1}{N} \sum_{j=1}^N r_u(\lambda_j) \quad \forall \quad i \quad (8)$$

where  $N$  is the total number of bands in the spectrum. The resulting  $r_u^*(\lambda_i)$  from the above equation is still unbiased,

where the variance for the difference of the measurement  $r_u^*(\lambda_i)$  and its true value  $\hat{r}_u^*(\lambda_i)$  is given by

$$\begin{aligned}
 \text{var}(r_u^*(\lambda_i) - \hat{r}_u^*(\lambda_i)) &= \text{var} \left( (r_u(\lambda_i) - \hat{r}_u(\lambda_i)) - \frac{1}{N} \sum_{j=1}^N (r_u(\lambda_j) - \hat{r}_u(\lambda_j)) \right) \\
 &= \text{var} \left( e_{\lambda_i} - \frac{1}{N} \sum_{j=1}^N e_{\lambda_j} \right) = \text{var} \left( \frac{N-1}{N} e_{\lambda_i} - \frac{1}{N} \sum_{j \neq i} e_{\lambda_j} \right) \\
 &= \frac{(N-1)^2}{N^2} \sigma^2 + \frac{N-1}{N^2} \sigma^2 = \frac{N-1}{N} \sigma^2
 \end{aligned} \tag{9}$$

Hence, we can see that by using the average value, the variances in the specular-free spectral values are reduced by half, being even slightly smaller than the variance of the measurement error in the original spectral values  $r_u(\lambda_i)$ , which equals  $\sigma^2$ .

Next, we aim at removing the shading factor in the specular free representation  $r_u^*(\lambda_i)$  in Equation 8. This is achieved by taking the band ratio values as follows.

$$\hat{F}_u(\lambda_i) = \frac{r_u^*(\lambda_i)}{\frac{1}{N} \sum_{j=1}^N |r_u^*(\lambda_j)|} + 1 \tag{10}$$

Here, to reduce the variance in the denominator, we have used the average value instead of the value of  $r_u^*(\lambda)$  at an arbitrary band since it has a lower variance. A bias value of 1 is added simply to make the feature values positive and does not change the nature of the representation. Consequently, the representation above has three advantages. Firstly, it does not depend on the geometry factor  $g(u)$ . Secondly, it has removed the specular component in Equation 5. Thirdly, it is quite resilient to measurement noise in individual bands. This is due to the subtraction of average band value in the computation of  $r_u^*(\lambda_i)$  defined in Equation 8. Moreover, we can use either form of  $r_u^*(\lambda_i)$  as presented in Equations 6 and 8 for the computation of  $\hat{F}_u(\lambda_i)$  in Equation 10. Since the use of Equation 6 leads to a higher variance than that yielded using Equation 8, we can expect Equation 8 to produce a lower representation error in Equation 10.

Nonetheless the corresponding theoretical analysis is intractable due to the normalisation with respect to the sum of random variables, this intuition is confirmed in practice. To this end, we illustrate how the treatment above affects the mean-squared error (MSE) for noise-corrupted spectra making use of two sample spectra. These are shown in the left-hand column of Figure 2. For each sample spectrum, we have perturbed its spectral values by adding zero-mean Gaussian noise with increasing variance. Feature vectors were computed for the noise-free and perturbed spectra based on the two different formulations mentioned above. The above steps were repeated 20 times for each value of noise variance. The average and standard deviations of the mean squared errors across different noise levels are shown in the plots on the right-hand panels of the figure. In the plots, the broken lines denote the error curves for the feature representation yielded by Equation 10 when Equation 6 is used. Note that, since Equation 6 assumes the use of a single arbitrary band  $\lambda_j$  for reference, in our error computations, and for the sake of comparison, we have randomly selected  $\lambda_j$  for each of the 20 trials for each value of noise variance. The continuous lines denote the error curves for the feature representation computed by substituting Equation 8 in

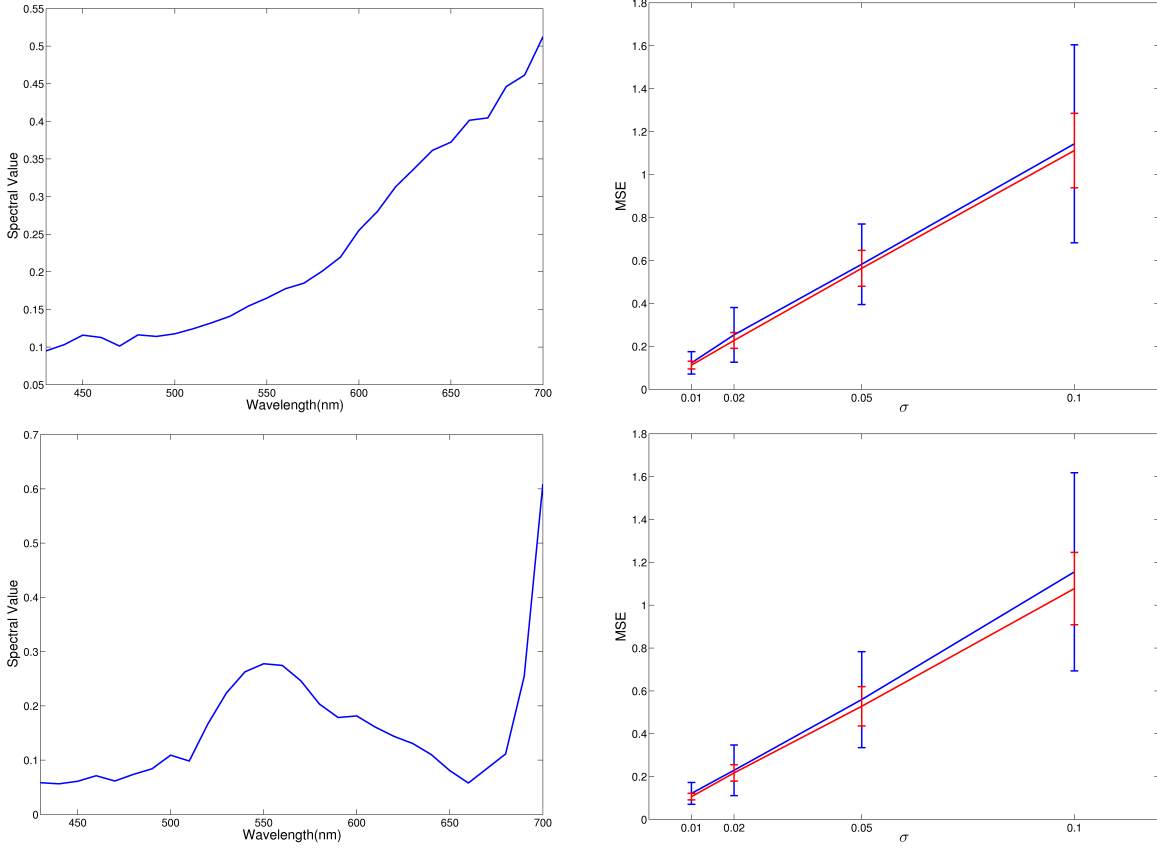


Fig. 2: Mean-squared error plots (right-hand panel) for two sample spectra in the left-hand column.

Equation 10. From the error plots, we can see that the average error values are not only smaller than those using a single, randomly selected band, but more importantly, have a much lower variance as compared with the alternative.

Also, note that the above representation does not explicitly take shadows into account. Shadows in indoor scenes are usually due to the influence of inter-reflection [37]. The illuminant energy incident upon pixel location  $u$  includes two components. The first of these arises from the illumination, i.e. the light source  $E(\lambda_i)$ . The second one is contributed by light reflected from other locations  $u$  in the scene is visible. We can express this using the quantity  $\sum_j q(u_j, u)I(\lambda_i, u_j)$ , where  $q(u_j, u)$  is the geometric factor determined by the relative angle between  $u_j$  and  $u$ . As a result, the radiance  $I(\lambda_i, u)$  at pixel location  $u$  can be modified by replacing  $E(\lambda_i)$  with  $E(\lambda_i) + \sum_j q(u_j, u)I(\lambda_i, u_j)$  in Equation 4. In this case, the subsequent derivation following the equation would not exactly remove the shadow component and, moreover, bias for the removal of shading and specular components may increase. There is no straightforward way to handle this problem. However, the derived representation in Equation 10 is still more robust to the influence of shadow effects than the raw spectral values. The reasons for this are twofold. Firstly, the inter-reflection term is usually quite small as compared to the light source energy  $E(\lambda_i)$ , making the Dichromatic model in Equation 4 a good approximation. Secondly, in the case that the pixel location  $u$  is foreshadowed by the light source so the only contribution is the inter-reflection one, the estimated  $r_u(\lambda_i)$  would be much smaller than the true albedo. Taking the ratio in Equation 10 would somehow balance this undesired effect. Similarly, for outdoor



shadows, which are usually caused by exposure to the skylight with a different colour temperature as opposed to the sunlight [38], the shadowed areas are also darker. This leads to a lower estimate of  $r_u(\lambda_i)$ . Taking the ratio would again balance this undesired effect. Moreover, if different light sources can be approximated by black body radiators, then the logarithm of band ratios are shadow invariant variables whose values are independent of the colour temperature as shown in [38]. Logarithms can be approximated as linear function when the ratio values are close to one due to the relation  $\log x \approx x - 1$  about  $x = 1$ .

So far, we have based our discussion on hyperspectral imaging in the terrestrial setting. We have also assumed that the light source spectrum  $E(\lambda)$  is known, so the reflectance spectra  $r_u(\lambda)$  can be obtained by normalising the image radiance value  $I(\lambda, u)$  with respect to  $E(\lambda)$ . In the terrestrial imaging setting, the light source spectrum  $E(\lambda)$  can be measured making use of a calibration target such as a white spectralon. Hence, we can make use of the reflectance spectra for material identification. The case for remote sensing imaging is very different though, where atmospheric conditions become the most important factor. There have been many references regarding atmospheric effects and physical assumptions in the literature [39], [40], [11], [41]. We do not deal with such effects explicitly in this paper as we are more interested in establishing a invariant representation to the photometric artifacts in the terrestrial setting. For remotely sensed imaging data, one can always use existing atmospheric correction methods [42], [43] to recover the scaled reflectance spectra. The invariant feature representation in Equation 10 can be employed to minimise the scaling artifacts. Alternatively, one can directly use the radiance spectra as the input to our algorithm. In this case, the proposed feature representation is no longer valid for specular removal, but still preserves shading invariance. Nevertheless, specular is much less of a problem for remote sensing as it is for terrestrial imaging. In summary, the proposed feature representation is effective in handling photometric artifacts in the terrestrial imaging setting and does not influence the performance in the remote sensing setting, where the assumptions made with respect to terrestrial imaging models may no longer hold.

#### IV. DISCRIMINANT ABSORPTION FEATURE SELECTION

It should be noted that, despite effective, the absorption band detection method in [35] can identify absorption bands that may not be significant for purposes of classification. For example, some absorption bands are due to the presence of water and, hence, arise in a wide variety of materials. Furthermore, the spectra of the light source may, potentially, skew the observed absorptions. These “spurious” absorptions must be eliminated before classification can be undertaken.

We cast this as a supervised feature selection problem. With all detected absorption bands at hand, only a few most discriminant bands are selected to form the final feature for classification. The other bands that do not contain sufficient discriminant information are not considered. Here we focus on the binary setting for feature selection, where training spectral samples are from two labelled classes. Extension to the multiclass problem is handled making use of a pairwise fusion framework and discussed in the next section. For absorption feature selection in a binary classification setting, the purpose is to identify those absorption bands for which the corresponding truncated

absorption spectral segments from positive and negative labelled classes are best separated. To this end, we propose a supervised algorithm based on Canonical Correlation Analysis (CCA) and the maximisation of the  $\alpha$ -divergence [44]. Our choice to the  $\alpha$ -divergence hinges in the notion that its use permits the measurement of the degree of dissimilarity of absorptions based upon the amount of information they provide to the classification tasks. Indeed, the  $\alpha$ -divergence is a generalisation of the Kullback-Leibler divergence and the Bhattacharyya coefficient [45]. It is based upon the Rényi entropy, which generalises Shannon's information measure. In this section, we commence by elaborating on the use of CCA as a means to subspace projection. We then show how the  $\alpha$ -divergence can be used as a basis for feature selection.

#### A. Canonical Correlation Analysis

For the supervised learning task, we are given a set of training spectra with known labels. The purpose is to infer a classifier so as to assign labels to the unknown testing spectra with minimum error. Let  $\mathcal{R} = \{r_u(\lambda)|u = 1, \dots, |\mathcal{R}|\}$  denote the set of training spectra and  $Z = \{z_u|u = 1, \dots, |\mathcal{R}|\}$  denote their corresponding labels, where  $r_u(\lambda)$  and  $z_u \in \{-1, 1\}$  represent the reflectance spectrum and the binary label for the  $u$ th training pixel respectively. We define  $\mathbf{F}_u^j = [\hat{F}_u(\lambda_1^j), \dots, \hat{F}_u(\lambda_2^j)]$  as the feature vector for the  $u$ th pixel and the  $j$ th absorption band identified by band edges  $(\lambda_1^j, \lambda_2^j)$ . The feature vector is computed from the truncated spectrum within the absorption band based on the representation scheme defined in Equation 10. Hence, each training spectrum has  $m$  feature vectors, where  $m$  is the total number of absorption bands recovered by the absorption detection algorithm discussed in the previous section. Our feature selection scheme requires that the pairs of feature vectors being compared have equal length. This is not necessarily the case for absorption features, since the length of the recovered absorptions may differ greatly. Moreover, evaluating the  $\alpha$ -divergence may be computationally expensive for very wide absorptions and can favor segments with higher dimensions. To this end, we use CCA as a preprocessing step to align the feature vectors so as to allow for feature selection. It aims at both, reducing the computational cost of the method and allowing the comparison of spectral segments of different lengths on an equal basis.

Let  $\mathcal{F}_i = \{\mathbf{F}_u^i|u = 1, \dots, |\mathcal{R}|\}$  denote the set of absorption features obtained from the  $i$ th absorption band in the training set. We perform CCA to discover the correlations between features extracted from any two absorption bands  $\mathcal{F}_i$  and  $\mathcal{F}_j$ . CCA recovers two bases vectors  $a$  and  $b$  for  $\mathcal{F}_i$  and  $\mathcal{F}_j$  respectively such that the one-dimensional projections of the feature vectors in the two sets along the direction of the corresponding basis are maximally correlated. Here, we have applied CCA to each pair of absorption feature vectors and recovered their respective bases. Hence, CCA is performed  $\binom{m}{2}$  times for all possible combinations of absorptions. Nevertheless, each CCA operation is similar to one another. Therefore in the following text, we use  $\mathcal{X}$  and  $\mathcal{Y}$  to represent feature sets  $\mathcal{F}_i$  and  $\mathcal{F}_j$  and denote the corresponding multi-dimensional random variables by  $X$  and  $Y$ . The problem for seeking

optimal projection directions can be formulated as follows

$$\begin{aligned}
\max_{a,b} \rho(X,Y) &= \frac{a^T \Sigma_{XY} b}{(a^T \Sigma_{XX} a)^{1/2} (b^T \Sigma_{YY} b)^{1/2}} \\
\mu_X &= E_X [\mathbf{x}] \quad \mu_Y = E_Y [\mathbf{y}] \\
\Sigma_{XX} &= E [(\mathbf{x} - \mu_X)(\mathbf{x} - \mu_X)^T] \\
\Sigma_{YY} &= E [(\mathbf{y} - \mu_Y)(\mathbf{y} - \mu_Y)^T] \\
\Sigma_{XY} &= E [(\mathbf{x} - \mu_X)(\mathbf{y} - \mu_Y)^T]
\end{aligned} \tag{11}$$

where  $E[\cdot]$  is the expectation operator,  $\mu_X$  and  $\mu_Y$  are the mean vectors for the random variables  $X$  and  $Y$ ,  $\Sigma_{XY}$  is the cross-covariance matrix between  $X$  and  $Y$  and  $\Sigma_{XX}$  and  $\Sigma_{YY}$  are the covariance matrices for  $X$  and  $Y$ , respectively. For the paired data sets  $\mathcal{X}$  and  $\mathcal{Y}$ , we can estimate the above terms using sample mean vectors and covariance matrices given by

$$\begin{aligned}
\hat{\mu}_{\mathcal{X}} &= \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i \quad \hat{\mu}_{\mathcal{Y}} = \frac{1}{N} \sum_{i=1}^N \mathbf{y}_i \\
\hat{\Sigma}_{\mathcal{X}\mathcal{X}} &= \frac{1}{N} \sum_{i=1}^N (\mathbf{x}_i - \hat{\mu}_{\mathcal{X}})(\mathbf{x}_i - \hat{\mu}_{\mathcal{X}})^T \\
\hat{\Sigma}_{\mathcal{Y}\mathcal{Y}} &= \frac{1}{N} \sum_{i=1}^N (\mathbf{y}_i - \hat{\mu}_{\mathcal{Y}})(\mathbf{y}_i - \hat{\mu}_{\mathcal{Y}})^T \\
\hat{\Sigma}_{\mathcal{X}\mathcal{Y}} &= \frac{1}{N} \sum_{i=1}^N (\mathbf{x}_i - \hat{\mu}_{\mathcal{X}})(\mathbf{y}_i - \hat{\mu}_{\mathcal{Y}})^T
\end{aligned} \tag{12}$$

where  $\mathbf{x}_i$  and  $\mathbf{y}_i$  denote the  $i$ th element in  $\mathcal{X}$  and  $\mathcal{Y}$  respectively, and  $N = |\mathcal{X}| = |\mathcal{Y}|$  is the total number of training samples in both data sets.

Let  $a' = \hat{\Sigma}_{\mathcal{X}\mathcal{X}}^{-1/2} a$  and  $b' = \hat{\Sigma}_{\mathcal{Y}\mathcal{Y}}^{-1/2} b$ . With this change of variables and the use of the plug-in estimators defined in 12, the CCA formulation in Equation 11 can be re-expressed as

$$\begin{aligned}
\max_{a',b'} \quad a'^T K b' &= a'^T \hat{\Sigma}_{\mathcal{X}\mathcal{X}}^{-1/2} \hat{\Sigma}_{\mathcal{X}\mathcal{Y}} \hat{\Sigma}_{\mathcal{Y}\mathcal{Y}}^{-1/2} b' \\
s.t. \quad a'^T a' &= 1 \quad \text{and} \quad b'^T b' = 1
\end{aligned} \tag{13}$$

which can be solved via singular value decomposition (SVD), where the maximum canonical correlation is given by the largest singular value of matrix  $K = \hat{\Sigma}_{\mathcal{X}\mathcal{X}}^{-1/2} \hat{\Sigma}_{\mathcal{X}\mathcal{Y}} \hat{\Sigma}_{\mathcal{Y}\mathcal{Y}}^{-1/2}$ , and the corresponding bases  $a'$  and  $b'$  are given by the left and right singular vectors corresponding to the largest singular value.

### B. Feature Selection via the $\alpha$ -Entropy

The previous CCA step provides subspace-projected scalar features for each pair of absorption features  $\mathcal{F}^i$  and  $\mathcal{F}^j$ . Let  $\mathbf{a}^T \mathcal{F}^i = \{\mathbf{a}^T \mathbf{F}_u | \mathbf{F}_u \in \mathcal{F}^i\}$  and  $\mathbf{b}^T \mathcal{F}^j = \{\mathbf{b}^T \mathbf{F}_u | \mathbf{F}_u \in \mathcal{F}^j\}$  denote the set of scalar feature values obtained by applying CCA for the  $i$ th and  $j$ th absorptions respectively, we further define  $\mathcal{X}^i = \{\mathbf{x}_u \in \mathbf{a}^T \mathcal{F}^i | z_u = 1\}$  for the subset of features from the  $i$ th absorption with positive labels and  $\mathcal{Y}^j = \{\mathbf{x}_u \in \mathbf{b}^T \mathcal{F}^j | z_u = -1\}$  for the subset

from the  $j$ th absorption with negative labels. The feature selection task in hand can be viewed as that of selecting those absorption bands with maximum discrimination between positive and negative spectral samples. This can be achieved by solving the following optimisation problem

$$\max_{\alpha} \{\pi\} = \max_{\alpha} \left\{ \sum_{i,j} D_{\alpha} (P(\mathcal{X}^i) || P(\mathcal{Y}^j)) \right\} \quad (14)$$

where  $P(\cdot)$  denotes the probability density function and  $D_{\alpha}(\cdot)$  refers to the divergence measure parameterised by  $\alpha$ . The divergence  $D_{\alpha}(\cdot)$  can be viewed as a measure of the separability between the two probability distributions  $P(\mathcal{X}^i)$  and  $P(\mathcal{Y}^j)$ . The purpose here is to find the parameter  $\alpha$  that maximises the total sum of pairwise separabilities.

Now we discuss the selection of the divergence measure  $D_{\alpha}$  in Equation 14. To this end, we make use of entropy, which is a measure of uncertainty for random variables. In information theory, Rényi [44] proposed the concept of  $\alpha$ -Entropy, which is a family of parametric functions defined over the random variable  $X$  with Probability Distribution Function (PDF)  $P(\mathcal{X}^i)$  parameterised by the scalar variable  $\alpha$  as follows

$$H_{\alpha}(\mathcal{X}^i) = \frac{1}{1-\alpha} \log \left( \int P(\mathcal{X}^i)^{\alpha} d\lambda \right) \quad (15)$$

Note that so far, and hereafter, we view  $\mathcal{X}^i$  and  $\mathcal{Y}^j$  as random variable with PDFs  $P(\mathcal{X}^i)$  and  $P(\mathcal{Y}^j)$ . This is important since we can characterise the dissimilarity between class conditional distributions of projected absorption features, where the divergence measures are usually used to model the separability between different classes. This has a number of advantages. Firstly, it permits the feature selection task to be cast as an optimisation setting as shown above. Secondly, we can select multiple bands simultaneously, recovering the optimal solution for the combination of absorptions under study. Thus, here, we make use of the Rényi  $\alpha$ -divergence [44] defined as

$$D_{\alpha}(P(\mathcal{X}^i) || P(\mathcal{Y}^j)) = \frac{1}{\alpha-1} \log \left( \int P(\mathcal{X}^i)^{\alpha} P(\mathcal{Y}^j)^{1-\alpha} d\lambda \right) \quad (16)$$

to recover the most discriminant absorption bands. We then obtain a specific form of the optimisation problem by implementing the  $\alpha$ -divergence  $D_{\alpha}(\cdot || \cdot)$  for each  $\mathcal{X}^i$  and  $\mathcal{Y}^j$  in Equation 14.

Here, we take a two step process to solve Equation 14 so as to recover the most discriminative features. Firstly, we recover the parameter  $\alpha$  for each pair so as to maximise the objective function. Secondly, we select the absorptions that yield the larger divergences given the recovered  $\alpha$ -parameters. To do this, we commence by using the shorthand  $\gamma = \frac{1}{\alpha-1}$  and writing the  $\alpha$ -divergence above as follows

$$D_{\alpha}(P(\mathcal{X}^i) || P(\mathcal{Y}^j)) = \gamma \log \left( \int P(\mathcal{X}^i) \left( \frac{P(\mathcal{X}^i)}{P(\mathcal{Y}^j)} \right)^{\frac{1}{\gamma}} d\lambda \right) \quad (17)$$

The change of variable above is important since it allows the algebraic manipulation necessary so as to use the lower bound of the  $\alpha$ -divergence for optimisation purposes. To this end, note that for  $\gamma > 0$ , we have  $\alpha > 1$ . Further, it is straightforward to see that the  $\alpha$ -divergence is monotonically increasing with respect to  $\alpha$ . As a result, by using the log-divergence, we can write

$$\mathcal{D}_{\alpha}(P(\mathcal{X}^i) || P(\mathcal{Y}^j)) = \log (D_{\alpha}(P(\mathcal{X}^i) || P(\mathcal{Y}^j))) = \log(\gamma) + \log \left( \log \left( \int P(\mathcal{X}^i) \left( \frac{P(\mathcal{X}^i)}{P(\mathcal{Y}^j)} \right)^{\frac{1}{\gamma}} d\lambda \right) \right) \quad (18)$$

Note that the first term on the left-hand side of the expression above does not depend on either of the random variables  $\mathcal{X}^i$  or  $\mathcal{Y}^j$ , i.e. our subspace projected absorptions. Thus, we remove it from further consideration and focus our attention on the second term. Since the logarithm is a monotonic function, maximising the log-divergence  $\mathcal{D}_\alpha(P(\mathcal{X}^i)||P(\mathcal{Y}^j))$  implies maximising

$$\hat{\mathcal{D}}_\alpha(P(\mathcal{X}^i)||P(\mathcal{Y}^j)) = \log \left( \int P(\mathcal{X}^i) \left( \frac{P(\mathcal{X}^i)}{P(\mathcal{Y}^j)} \right)^{\frac{1}{\gamma}} d\lambda \right) \quad (19)$$

To turn the problem of optimising over the log-divergence above into one that can be solved algebraically, we use the following relation presented in [46]

$$\int P(\mathcal{X}^i) \log \left( \frac{P(\mathcal{X}^i)}{P(\mathcal{Y}^j)} \right)^{\frac{1}{\gamma}} d\lambda = \frac{1}{\gamma} KL(P(\mathcal{X}^i)||P(\mathcal{Y}^j)) \leq \log \left( \int P(\mathcal{X}^i) \left( \frac{P(\mathcal{X}^i)}{P(\mathcal{Y}^j)} \right)^{\frac{1}{\gamma}} d\lambda \right) \quad (20)$$

where  $KL(P(\mathcal{X}^i)||P(\mathcal{Y}^j))$  is the Kullback-Leibler (KL) divergence between the subspace-projected absorptions given by

$$KL(P(\mathcal{X}^i)||P(\mathcal{Y}^j)) = \int P(\mathcal{X}^i) \log \frac{P(\mathcal{X}^i)}{P(\mathcal{Y}^j)} d\lambda \quad (21)$$

The expression above is important since it provides a lower bound for  $\hat{\mathcal{D}}_\alpha(P(\mathcal{X}^i)||P(\mathcal{Y}^j))$ . This permits using the lower bound of the log-divergence as an alternative to  $\hat{\mathcal{D}}_\alpha(P(\mathcal{X}^i)||P(\mathcal{Y}^j))$ . Further, note that the KL divergence is, in general, not symmetric, i.e.  $KL(P(\mathcal{X}^i)||P(\mathcal{Y}^j)) \neq KL(P(\mathcal{Y}^j)||P(\mathcal{X}^i))$ . To obtain a symmetric log-divergence, we can employ the average distance over the two possible KL divergences for the two distributions. This yields the Jeffreys Divergence [47], which can be viewed as an expected value over the KL divergences for those random variables whose distributions are exponential in nature.

Thus, we can use, as an alternative to  $\pi$  in Equation 14, the quantity

$$\hat{\pi} = \sum_{i,j \in \Gamma} \frac{1}{2\gamma} \left( KL((P(\mathcal{X}^i)||P(\mathcal{Y}^j))) + KL((P(\mathcal{Y}^j)||P(\mathcal{X}^i))) \right) = \sum_{i,j \in \Gamma} \frac{1}{\gamma} \mathcal{KL}((P(\mathcal{X}^i), P(\mathcal{Y}^j))) \quad (22)$$

where  $\mathcal{KL}((P(\mathcal{X}^i), P(\mathcal{Y}^j))) = \frac{1}{2} \left( KL((P(\mathcal{X}^i)||P(\mathcal{Y}^j))) + KL((P(\mathcal{Y}^j)||P(\mathcal{X}^i))) \right)$  is the Jeffreys divergence, i.e. the symmetrised Kullback-Leibler (KL) divergence. The summation is taken over the set of all pairs  $\Gamma = \{(i, j) | i, j \in [1, \dots, m] \text{ and } i < j\}$  with  $m$  being the total number of absorption bands detected.

Moreover, we can write the expression above in compact form. We do this by constructing a matrix  $\mathbf{D}$  of order  $|\Gamma|$  whose element  $D_{i,j}$  indexed  $i, j$  is given by the symmetrised KL divergence between every pair of subspace-projected absorption segments, i.e.  $D_{i,j} = \mathcal{KL}((P(\mathcal{X}^i), P(\mathcal{Y}^j)))$ . Thus,  $D_{i,j}$  is defined as the divergence between the positively labelled subset of subspace projected features for the  $i$ th absorption band and negatively labelled subset of projected features for the  $j$ th absorption band, as delivered at output by the CCA processing.

By expressing  $\frac{1}{\gamma}$  as the product of two other variables  $\phi_i, \phi_j$  for each random-variable pair  $\mathcal{X}^i, \mathcal{Y}^j$  in Equation 22, we can reformulate the optimisation problem defined in Equation 14 as follows

$$\max\{\hat{\pi}\} = \max_{\phi} \left\{ \sum_{i,j} \phi_i D_{i,j} \phi_j \right\} = \max_{\phi} \{ \phi^T \mathbf{D} \phi \} \quad (23)$$

From inspection, it is clear that maximising  $\hat{\pi}$  can be posed as an eigenvalue problem, and that  $\phi$  is an eigenvector of  $\mathbf{D}$  whose corresponding eigenvalue is  $\hat{\pi}$ . Furthermore, the expression above is reminiscent of the Rayleigh quotient. Note that the elements of the vector  $\phi$  are required to be real and positive. This is a consequence of the fact that  $\gamma = \frac{1}{\alpha-1}$  and that  $\alpha$  is bounded between  $-1$  and unity. Consequently, the coefficients of the eigenvector  $\phi$  are always non-negative. Since the elements of the matrix  $\mathbf{D}$  are non-negative, it follows that the quantity  $\phi^T \mathbf{D} \phi$  should be positive. Hence, the set of solutions reduces itself to those that correspond to a constant  $\hat{\pi} > 0$ . We also require the coefficients of the vector  $\phi$  to be linearly independent of the all-ones vector  $\mathbf{e} = [1, 1, \dots, 1]^T$ .

With these observations in mind, we focus on proving the existence of a vector  $\phi$ , linearly independent from  $\mathbf{e}$ , and demonstrating that this vector is unique. To this end, we use the Perron-Frobenius theorem [48]. This concerns the proof of existence regarding the eigenvalue  $\hat{\pi} = \max_{i=1,2,\dots,|\Gamma|} \{\pi_i\}$  of a primitive, real, non-negative, symmetric matrix  $\mathbf{D}$ , and the uniqueness of the corresponding eigenvector  $\phi$ . The Perron-Frobenius theorem states that the eigenvalue  $\hat{\pi} > 0$  has multiplicity one. Moreover, the coefficients of the corresponding eigenvector  $\phi$  are all positive and the eigenvector is unique. As a result, the remaining eigenvectors of  $W$  have at least one negative coefficient and one positive coefficient. If  $\mathbf{D}$  is substochastic,  $\phi$  is also known to be linearly independent of the all-ones vector  $\mathbf{e}$ . As a result, the leading eigenvector of  $\mathbf{D}$  is the unique solution of Equation 23.

Thus, we can recover the optimum  $\alpha$  for each pair of subspace-projected absorptions  $\mathcal{X}^i$  and  $\mathcal{Y}^j$  by building the matrix  $\mathbf{D}$  and computing its leading eigenvector  $\phi$ . Once the eigenvector  $\phi$  is at hand, we can recover the  $\alpha$  parameter by using the relationship  $\alpha = 1 + \frac{1}{\gamma} = 1 + \phi_i \phi_j$  presented earlier. Once the  $\alpha$  parameter has been recovered, those pairs  $X, Y$  corresponding to the largest divergences  $\hat{\mathcal{D}}_\alpha(P(\mathcal{X}^i) \| P(\mathcal{Y}^j))$  can be selected for purposes of classification.

## V. ALGORITHM

With the theoretical background of our method above, we now elaborate further on our algorithm for material identification. The step-sequence of the algorithm is presented in Figure 3. The idea underpinning the algorithm here is that the optimal absorption band should best discriminate between positive and negative spectral classes. This can be achieved by maximising the  $\alpha$ -divergence between distributions of corresponding spectral features.

Note that, for every candidate absorption band, each spectral segment can be seen as a feature vector in  $N$ -dimensional space, where the dimensionality  $N$  is given by the number of bands in the absorption spectral segment. As the number of bands may be different for every pair of candidate absorptions, the computed KL divergence may be biased towards those features with a greater number of bands. This is regardless of the separability of the two classes and is due to the fact that distributions in higher dimensional spaces are, in general, more scattered than those defined in lower dimensions. To balance this effect, we apply CCA to each of pair of candidate absorptions as a preprocessing step for the algorithm above. Note that, since we have at our disposal training data, the sample covariance and cross-covariance matrices used by the CCA can be computed directly. Thus, CCA will project all the spectral segments for every candidate absorption band onto a one-dimensional space. We then compute

Given the set of training spectral features  $\mathcal{R} = \{r_1(\lambda), \dots, r_{|\mathcal{R}|}(\lambda)\}$  and the label set  $\mathcal{Z} = \{z_u\}$ , where  $z_u \in \{-1, 1\}$  is the label for the spectrum  $r_u(\lambda)$ , we compute the set of tuples  $\{(\lambda_1^1, \lambda_2^1), \dots, (\lambda_1^m, \lambda_2^m)\}$  corresponding to the extreme spectral values for the  $m$  absorption bands recovered by our algorithm, as described in the previous section. For each tuple  $(\lambda_1^j, \lambda_2^j)$ , do

- Extract the spectral bands for the recovered absorption segments and transform them to invariant feature representation  $\hat{F}_u(\lambda_i)$  as shown in Section 3.2.
- For each pair of absorption features, apply CCA to the feature vectors and project them along the direction given by the maximum canonical correlation.
- Estimate the symmetrised KL divergence values for the projected feature elements via Equation 24.
- With the symmetrised KL divergences, construct the matrix  $\mathbf{D}$  and recover its leading eigenpair. Use the corresponding eigenvector to recover the values of  $\alpha$  that maximise the divergence between classes. Use the  $n$  absorption bands which correspond to the  $n$ -largest  $\alpha$ -divergence values.

Fig. 3: Discriminant Absorption Band Selection Algorithm

the KL divergence for the distributions of these subspace-projected, one dimensional features and select the most discriminative absorptions.

Since CCA is a correlation based technique, in the algorithm above we have assumed the positive and negative sample-set distributions to be Gaussian in nature. This is a reasonable assumption due to the unimodality of the absorptions. Since linear projection preserves Gaussian distributions, the subspace-projected feature elements also satisfy Gaussian distributions whose KL divergence can be computed in closed form. Let  $\hat{\mu}_{\mathcal{X}}$  and  $\hat{\sigma}_{\mathcal{X}}^2$  be the sample mean and variance of the subspace-projected feature set  $\mathcal{X}$ , and  $\hat{\mu}_{\mathcal{Y}}$  and  $\hat{\sigma}_{\mathcal{Y}}^2$  be the sample mean and variance of  $\mathcal{Y}$ , the symmetrised KL divergence between distributions  $P(\mathcal{X})$  and  $P(\mathcal{Y})$  is then given by

$$\mathcal{KL}(P(\mathcal{X})||P(\mathcal{Y})) = \frac{(\hat{\mu}_{\mathcal{X}} - \hat{\mu}_{\mathcal{Y}})^2(\hat{\sigma}_{\mathcal{X}}^2 + \hat{\sigma}_{\mathcal{Y}}^2)}{2\hat{\sigma}_{\mathcal{X}}^2\hat{\sigma}_{\mathcal{Y}}^2} + \frac{\hat{\sigma}_{\mathcal{X}}^2}{2\hat{\sigma}_{\mathcal{Y}}^2} + \frac{\hat{\sigma}_{\mathcal{Y}}^2}{2\hat{\sigma}_{\mathcal{X}}^2} - 1 \quad (24)$$

In the expression above, for brevity, we have omitted the superscripts for the subspace-projected features  $\mathcal{X}$  and  $\mathcal{Y}$  without further disambiguation. The KL divergences estimated from the above equation are then used to compute the matrix elements  $D_{i,j}$  in Equation 23, where  $D_{i,j} = \mathcal{KL}(P(\mathcal{X}^i), P(\mathcal{Y}^j))$  captures the divergence between positively labelled subspace projected features from the  $i$ th absorption band  $\mathcal{X}^i$  and negatively labelled projected features from the  $j$ th absorption band  $\mathcal{Y}^j$ .

Finally, we have employed a Support Vector Machine (SVM) [49] for the classification of absorption features extracted by our algorithm. For  $n$  selected absorption bands, we take the truncated absorption spectral segment from each absorption band making use of the feature representation proposed in Section III. The truncated absorption feature segments, scaled by their corresponding eigenvalues of  $D$  in Equation 23, are then concatenated to form the final absorption feature vector to be used for SVM classification. To solve multiclass classification problems, we

have adopted a pairwise fusion framework [50]. The idea is to build a binary classifier for any pair of two classes, where the final classification result is obtained by combining the output of all binary classifiers. Instead of voting on hard decisions, we aggregate the decision value output produced by each component classifier and assign the testing data to the class with highest aggregation value. This can be viewed as a generalisation to majority voting. The procedure of the fusion algorithm is presented in Figure 4.

---

### Training

Given a set  $\Gamma = \{\omega_1, \omega_2, \dots, \omega_{|\Omega|}\}$  of materials to identify

- Obtain positive samples for every pair of materials  $\omega_i$  and  $\omega_j$  in  $\Omega$ .
- Perform absorption feature extraction and selection for every pair of classes  $\omega_i$  and  $\omega_j$ .
- Train the SVM classifier over the selected features.

### Testing

- For any novel test spectrum, apply each of the trained binary classifiers above
- Combine the binary classification results and assign the spectrum to the class that maximises the total posterior probability given by

$$\xi = \arg \max_i \sum_{j=1, j \neq i}^N Q_{i,j} \quad (25)$$

where  $Q_{i,j}$  is the decision value output of the classifier for the pair of classes indexed  $i$  and  $j$  in  $\Omega$ .  $Q_{i,j}$  is anti-symmetric in that  $Q_{i,j} = -Q_{j,i}$  for any  $i \neq j$ .

---

Fig. 4: Pairwise classifier fusion method

The use of a pairwise classifier fusion framework has two major advantages over a single classifier directly applied to the multiclass classification task. Firstly, it is often easier to separate two classes and combine the binary classification results than finding a multiclass classifier that can distinguish all classes. Secondly and, more importantly, the use of pairwise fusion framework allows us to use different sets of absorption features for every pair of classes. In our case, we can use those absorption bands which best distinguish between different classes without being constrained to a unique band or feature that is applied to all material classes under study.

It is worth stressing in passing that, despite our focus is on classification problems, the proposed pairwise classification method can also be used for the more complicated problem of supervised spectral unmixing [31]. In classification, each pixel in the image is assumed to have a pure spectrum and assigned to a single material class. This may not be the case for remote sensing image data, where image pixel spectra may correspond to multiple material types due to the variability in material composition and restriction in spatial resolution. To address this problem, we note that the SVM classifier can also produce probability values at output for each label class [51]. We can exploit these probabilistic output values from each binary classifier and aggregate them via the right-hand



side of Equation 25, which is the final probability for class  $i$  (up to a scaling factor) corresponding to the mixing coefficient of  $i$ th material class. The only difference is that now each  $Q_{i,j}$  is a probability value and  $Q_{j,i} = 1 - Q_{i,j}$ .

## VI. EXPERIMENTAL RESULTS

Having presented the theoretical foundation for our method, in this section we perform experiments so as to demonstrate the performance of our algorithm on both remotely sensed and terrestrial hyperspectral imagery. In the first experiment, we apply our algorithm to the AVIRIS hyperspectral imagery of NW Indiana Pines that has been widely used as a benchmark data set for hyperspectral image classification in previous studies [2], [19], [20], [21], [22]. The image consists of  $145 \times 145$  pixels with 220 spectral bands covering the visible and near infrared spectral range from 400nm to 2200nm with fine spectral resolution, where 20 out of those 220 spectral bands have been identified as water absorption bands. For experiments on ground-based imagery, we have studied classification of wheat infected with Stem Rust (*Puccinia Striformis*) and imagery of assorted nuts. We have collected our imagery using a hyperspectral camera system from Opto-Knowledge Systems Inc (OKSI) [52], which is comprised by a broadband monochrome camera and a liquid crystal tunable filter (LCTF). The system comprises a camera pair, one of which operates in the visible (VIS) spectrum and the other one in the near-infrared (NIR) range. The spectral resolution of the VIS LCTF is 10nm between the 400nm and 720nm wavelengths, i.e. 33 bands covering the visible spectrum. Similarly, the spectral resolution of the NIR LCTF is 10nm between the 650nm to 1100nm wavelengths, i.e. 46 bands covering the near infrared spectrum. Its worth noting that, in our experiments, we use the VIS LCTF for wheat imagery, and both, the VIS and NIR filters for the images for the nuts. For both, the VIS and the NIR imagery, the spatial resolution is 696 by 520 pixels. The light source spectrum was measured using a white spectralon target. Radiometric calibration is performed by normalising pixel spectral values with respect to the light source spectrum in a bandwise manner.

For our classification experiments, we have compared the proposed absorption learning approach with the alternative of applying a SVM classifier directly to the full spectra. For remote sensing data, we have also tested with another alternative SVM-based classification method by applying the classifier to the spectra, preprocessed so as to remove the water absorption bands. We used the LIBSVM package [53] for training RBF kernel SVM classifiers for all approaches. Each input feature is linearly rescaled to the range of  $[0, 1]$  before SVM training to accommodate for different dynamic ranges and, hence, the input variables and the testing data are scaled accordingly based on the training data. The SVM parameters were chosen via 5-fold cross validation on the training data set. Each classification test has been repeated 5 times with different data pools of randomly selected training and testing sets. We have recorded the classification accuracy rate, i.e. the percentage of correctly classified testing samples in the testing set, obtained in each round and calculated the mean value and standard deviation of the accuracy rates over different rounds.

It is also worth noting that our proposed discriminant absorption learning method is quite efficient in testing. Given a testing example, the absorption spectral segments corresponding to the selected absorption bands are extracted.

Then the photometric invariant transforms are applied to the truncated segments to form the feature vector. Finally the learned SVM classifier is applied to the feature vector. It does not involve any absorption detection at the testing stage. The only overhead over the direct application of SVM to the raw spectra is in the photometric invariant transform, which only involves simple algebraic operations. The training stage takes a bit more time, as absorption detection needs to be performed on each training spectra. Additional overheads are introduced by performing CCA, computing the KL divergence and the eigen-decomposition. However, the computational cost of these steps is still cheaper than training the SVM on the raw spectra. For the data sets studied in our experiment, training takes from a few seconds to a couple of minutes, depending on the training data size, where most of the time is spent on the SVM training.

#### A. Results on Remotely Sensed Hyperspectral Imagery

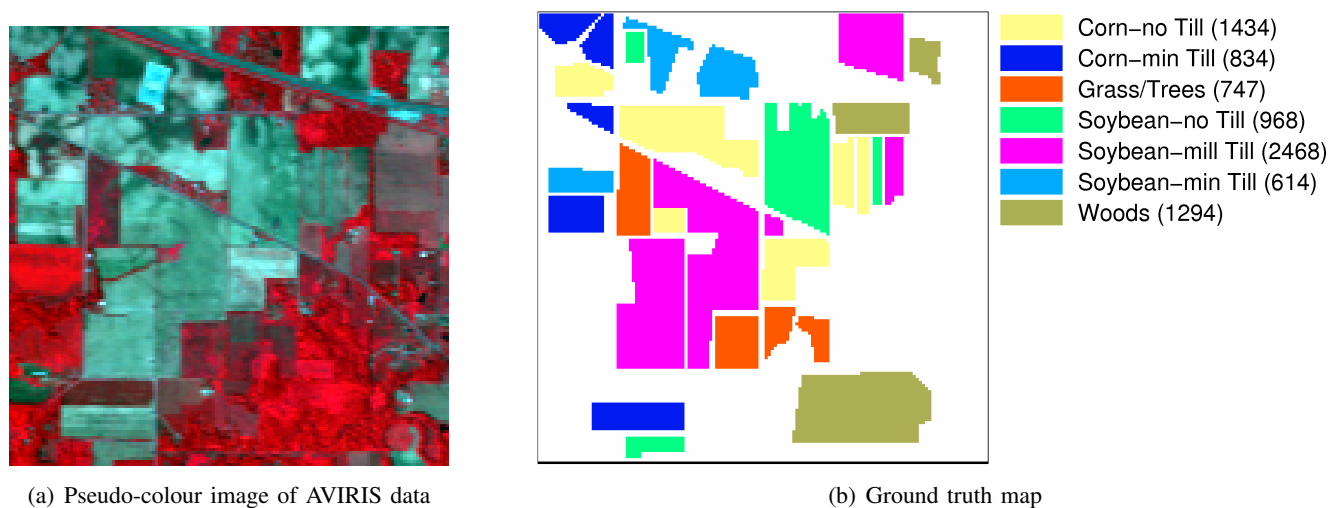


Fig. 5: AVIRIS data set of NW Indiana Pines.

The AVIRIS hyperspectral imagery data set of Indiana Pines contains 16 ground-truth classes with varying sizes from 20 to 2468 pixels per class. We have chosen the 7 classes with maximum numbers of labelled pixels so as to guarantee that there is enough labelled data for the classes under consideration. The synthesised pseudo-colour image of the hyperspectral imagery as well as the ground truth for the 7 classes under study are shown in Figure 5. Note that different land covers are represented by different colours in the ground truth map and the number following each legend denotes the size of each class. We have examined the classification performance with 5%, 10% and 25% of the labelled pixels for training and the rest for testing for the three methods under consideration, i.e. our proposed absorption learning method (ABS+SVM), direct application of SVM to raw spectra (SVM1) and the application of the SVM to the preprocessed spectra with the water absorption bands removed (SVM2). For the AVIRIS imagery data, the absorption features are mostly associated with water absorptions and thus not discriminative for purposes of classification. To this end, we have augmented the set of absorption features with

their complementary features for feature selection. Each complementary feature vector contains all spectral values except those taken out from the identified absorption band.

The mean accuracy rates and standard deviations of our approach and the alternatives are reported in Table I. From the table, we can see that our ABS+SVM method consistently achieves the highest performance regardless of the size of the set used for training. It gains a margin of advantage over SVM2 applied to the clean spectra and performs much better than SVM1 applied to raw noisy spectra. Both ABS+SVM and SVM2 achieve lower variance in classification accuracy as compared to SVM1. This implies that they are both more stable with respect to the random selection of training data. However, ABS+SVM is applied to the raw noisy spectra in the same setting as SVM1. In contrast, SVM2 is applied to the clean data with noisy bands removed. Hence we can see that ABS+SVM is quite robust to noise corruption and outliers when applied directly to the raw spectra. Moreover, expert knowledge is needed to identify the water absorption bands and remove them manually before the application of SVM2. Note that, in the case of ABS+SVM, the method can be applied automatically.

	SVM1	SVM2	<b>ABS+SVM</b>
5%	$73.47 \pm 1.39$	$82.52 \pm 0.77$	$83.84 \pm 0.74$
10%	$80.14 \pm 1.05$	$87.09 \pm 0.55$	$88.26 \pm 0.34$
25%	$86.50 \pm 0.56$	$91.21 \pm 0.37$	$92.09 \pm 0.41$

TABLE I: Performance comparison of our approach (ABS+SVM) vs. the alternatives

A closer look at the bands selected further justifies the validity of the proposed learning approach and the negative influence of water absorptions on classification. Since the randomness in training and testing, different absorption bands or their complement might be chosen for different pairs of classes within each different training set. We have recorded the top three selections and all of them are complementary features. The removed absorption bands are shown in Figure 6 along with images corresponding to the removed bands. It can be seen that two of three absorption (1 and 3) bands are from water absorptions, which do not contribute to classification. The second band removed corresponds to the low contrast of different land covers, which is also unlikely to carry much discriminant information and, hence, suppressed by our method.

Next, we examine the features used by our ABS+SVM algorithm for classification. Since our algorithm selects features of different lengths for each pairwise SVM classifier, we have applied Kernel principal component analysis (KPCA) [6] to the kernel matrix induced by the feature vectors and kept the first two principal components for visualisation purposes. The distributions of the KPCA projections for the absorption features extracted for each pair of classes with 10% of labelled pixels as training data are plotted in the right-hand column of Figure 13. For purposes of comparison, we also plot the KPCA embeddings for the spectral features used by SVM1 on the left-hand column of the figure. For the sake of brevity, we only show a selected number of class pairs with noticeable differences. For each subfigure, the blue dots represent feature space projections for the first class, and the green dots represent those for the second class. From the feature projections, we can appreciate that the invariant absorption

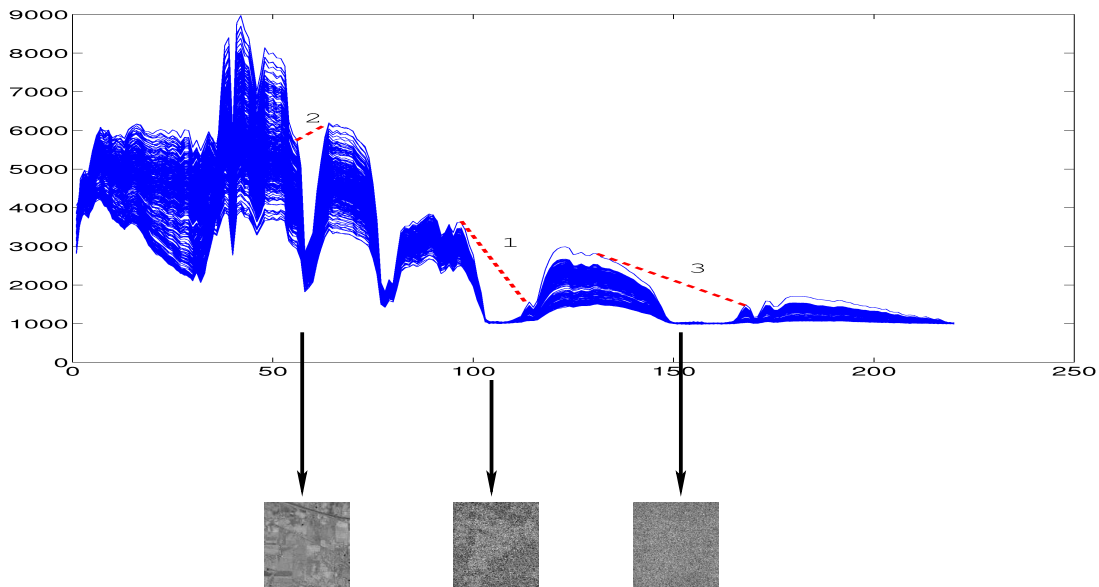


Fig. 6: Top three absorption bands identified for removal by our method for classification on the AVIRIS data.

features bear more robustness than the preprocessed spectra with noisy water absorption bands filtered out. We can see that different coloured dots are well separated. In contrast, in the embeddings of filtered spectra the existence of outliers is more evident.

Finally, in Figure 8, we show the surface maps delivered by the three classification methods applying the learned classifier models to the foreground pixels of the image. The left-most, middle and right-most columns in the figure correspond to surface maps recovered by SVM1, SVM2, and ABS+SVM, respectively. It can be seen that the surface maps delivered by SVM2 and the proposed ABS+SVM are in better accordance with the ground truth in Figure 5(b). In turn ABS+SVM delivers slightly more consistent surface maps than the one yielded by SVM2. This is especially true for small training sample sizes.

### B. Results on Terrestrial Hyperspectral Imagery

We now turn our attention to material classification terrestrial hyperspectral imagery. This is a more challenging classification task as compared to remote sensing applications. This is as the former is often complicated by the photometric conditions in the scene due to surface geometry, which introduces further intra-class variations for the same material.

*1) Results on the Wheat:* We focus on wheat infected by stem rust for our first experiments on terrestrial imagery. Two hyperspectral images were taken for two specimens of wheat, one healthy and one infected. These are illuminated from two different light source directions. The pseudo-colour images are shown in the top row of Figure 9. In both panels, the healthy wheat was placed in the left-hand side and the infected wheat on the right-hand side of the scene. The two images differ only in terms of light source direction. For the image in Figure 9(a), the illuminant was placed approximately  $45^\circ$  left of the camera axis. In Figure 9(b), the light source direction is

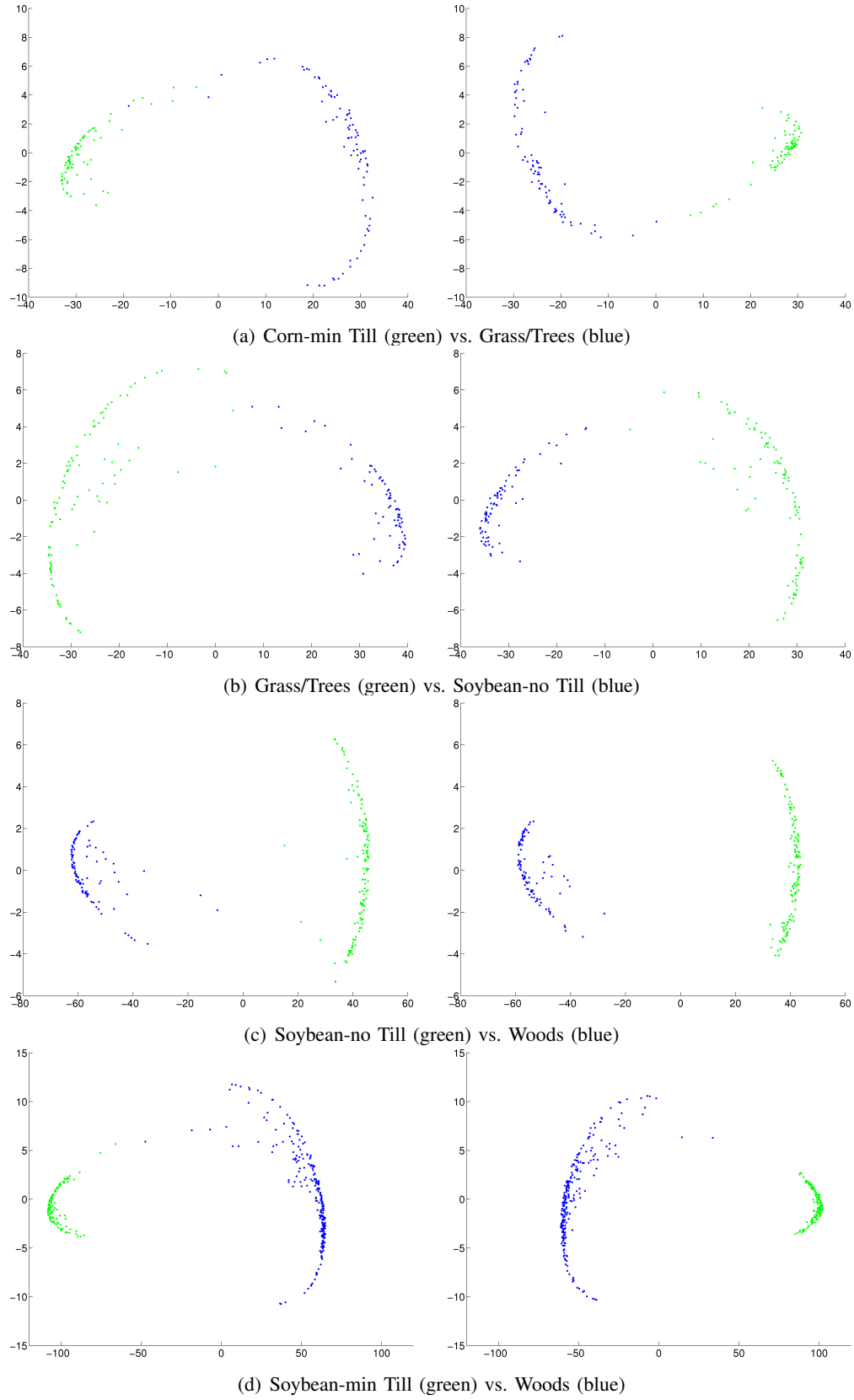


Fig. 7: KPCA mappings for selected pairs of land covers used in our experiments. Left-hand column: KPCA embeddings for the spectral features used by SVM1; Right-hand column: KPCA embeddings for the features selected by the ABS+SVM method.



Fig. 8: Surface maps yielded by our approach (ABS+SVM in the rightmost column) and the alternatives (SVM1 in the leftmost column and SVM2 in the middle column) for different training data sizes. Legends are consistent with the ground truth map shown in Figure 5(b).

approximately  $45^\circ$  right of the camera direction. The healthy and infected plants look very similar and is hard to distinguish them based upon pixel colour. We randomly picked 100 pixels from both, the healthy and the infected wheat from the first image in Figure 9(a). The plots for the spectra of these pixels are shown in Figure 9(c). From the plots, it is clear that, despite scale variations due to pixel intensity, the spectral traces for both specimens have similar shapes. We applied our discriminant band selection algorithm, which selects a single most discriminative band segment corresponding to the absorption between 550nm and 720nm, as indicated by the blue arrows in Figure 9(c). The plots for the invariant absorption features recovered by our algorithm are shown in Figure 9(d), where green curves denote features of healthy wheat and red traces correspond to infected tissue.

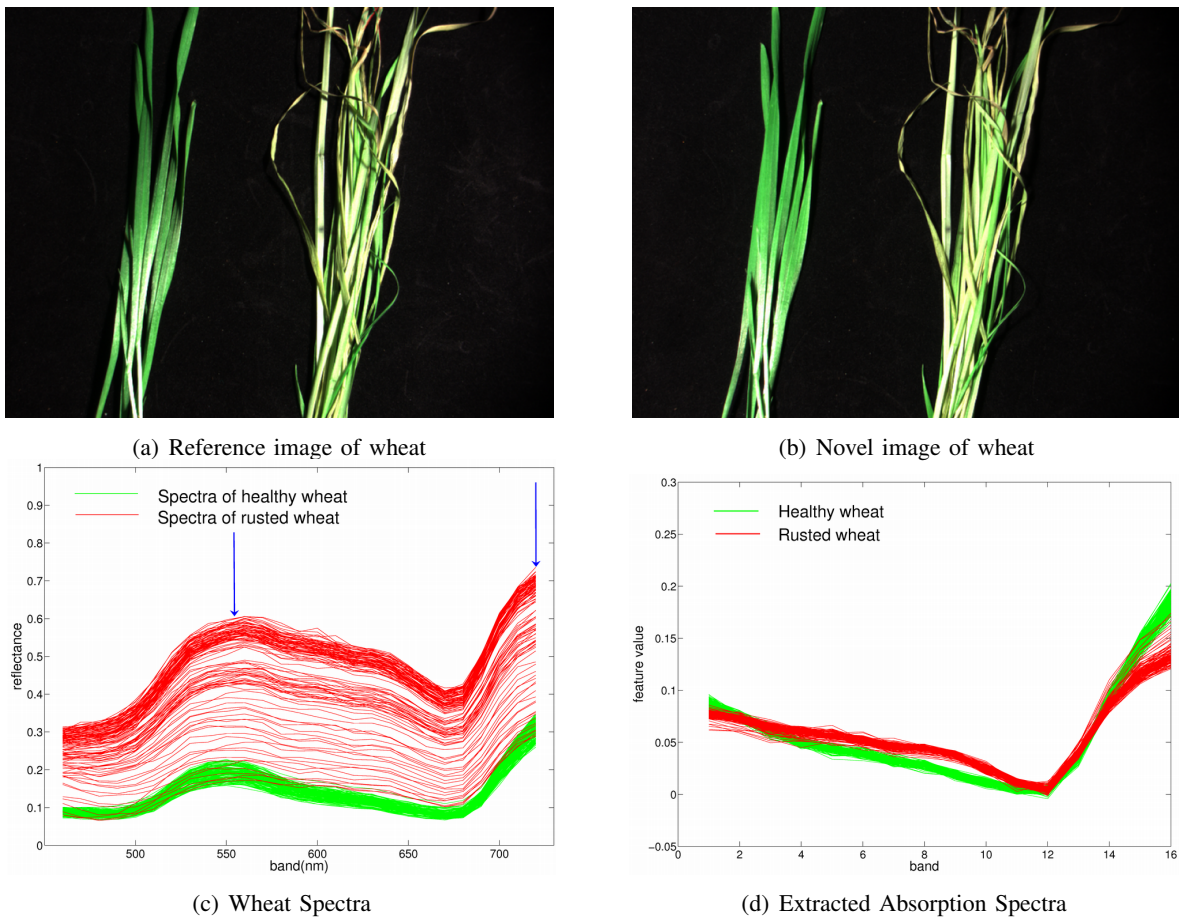


Fig. 9: Pseudo-colour imagery of wheat and corresponding absorption spectra.

To take our analysis further, we now provide a quantitative analysis on the classification performance. Specifically, we have used the image in Figure 9(a) as the reference one where training spectra are selected. The image in Figure 9(b) is the novel testing image. We have then trained the classifiers using both, the absorption features delivered by our algorithm and the raw spectra and tested the SVMs on both images. The alternative is, hence, equivalent to SVM1 in the remotely sensed data experiments. Here, for the sake of brevity, we denote the alternative SVM. In Figure 10, we show sample surface maps rendered by our algorithm and the alternative using a training sample size of 100 spectra, i.e. pixels, for each class. The surface maps recovered for the reference image as shown in the top panel of Figure 10, are quite similar for the two algorithms. Nonetheless, for the novel image, the surface map recovered by our algorithm is better than that delivered by the alternative, as can be seen from the bottom panel of Figure 10.

Note that, in our experiments, the number of training pixels chosen from each class will clearly impact in the results yielded by both methods. In Table II, we show the classification accuracy rates and standard deviations for our algorithm (ABS+SVM) and the alternative (SVM) as a function of training-set size. Note that classification accuracies for the reference image were calculated with the selected training pixels excluded. From Table II, we can conclude that our method is more stable, performance-wise, than the direct application of SVMs to the raw

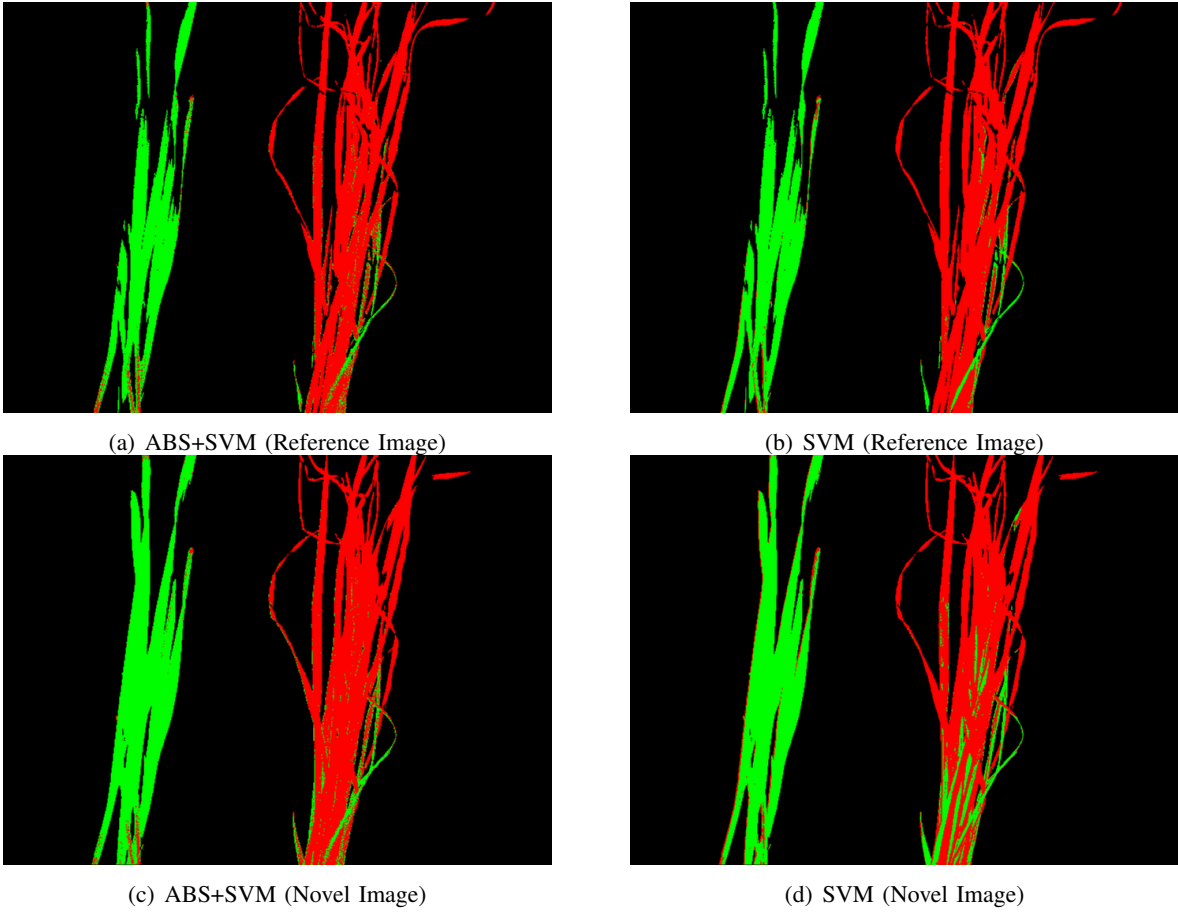


Fig. 10: Results yielded by our method and the alternative. Left-hand column: Surface maps delivered by ABS+SVM; Right-hand column: Results yielded by the SVM.

TABLE II: Accuracy rates and standard deviations for our method (ABS+SVM) and the alternative (SVM) on the wheat data

Sample size	Error rate(%)			
	Reference Image		Novel Image	
	SVM	<b>ABS+SVM</b>	SVM	<b>ABS+SVM</b>
10	80.65 $\pm$ 1.95	83.19 $\pm$ 1.91	74.61 $\pm$ 3.17	82.59 $\pm$ 1.76
50	92.72 $\pm$ 1.37	92.88 $\pm$ 1.10	89.73 $\pm$ 1.60	93.01 $\pm$ 1.16
100	95.22 $\pm$ 0.58	94.83 $\pm$ 0.84	90.82 $\pm$ 2.19	94.38 $\pm$ 0.67
500	96.50 $\pm$ 0.52	96.21 $\pm$ 0.51	92.32 $\pm$ 1.54	95.04 $\pm$ 0.34
1000	96.90 $\pm$ 0.37	96.70 $\pm$ 0.22	93.18 $\pm$ 1.95	95.37 $\pm$ 0.60



spectra. Firstly, while blind SVM training can do quite a good job on the reference image from which the training samples were selected, it fails to classify as accurately as our absorption-based algorithm the data on the novel testing image in Figure 9(b). This problem can be somehow alleviated by using more training samples, but in the case of our method, no matter how many training samples are used, performance on the novel image is always better than that delivered by the alternative. Secondly, the variation of the test results for the raw spectral data is larger than that for our algorithm. As training samples are chosen randomly, this indicates that the alternative relies more on the quality of the training samples, i.e. it requires the use of clean data for training. In contrast, our algorithm is more robust to noise and, hence, is less dependent on the choice of the training sample-set. Thirdly, the alternative requires a reasonably large sample set for training. For small training sample sizes, ABS+SVM consistently outperforms the alternative on both images.

2) *Results on Nuts:* We now turn our attention to imagery of assorted nuts and compare the classification performance. To this end, two pairs of test images are taken under incandescent light at different lighting and viewing angles. Each image pair contains both, the VIS and the NIR views. The two images in each pair capture the same scene depicting a number of nut-types. A composite image comprising all the bands covering both, the visible and NIR spectral ranges is synthesised through image fusion [54]. This is achieved by registering the two images and concatenating the VIS and NIR spectra at corresponding pixel-locations. Image registration is performed by manually selecting corresponding control points in both images and recovering the projective transformation from the NIR view to the visible image such that the average reprojection error over the set of control points is minimised.

Our analysis departs from the two full-range hyperspectral images, i.e. between 380 and 1100nm in 10nm steps, of assorted nuts. Figure 11 shows the pseudo-colour images for the two hyperspectral images under study along with the example spectra of nuts being classified. Since we are interested in the classification task, the background has been removed using a preprocessing step based upon thresholding and morphological operations. In our imagery, there are five nut types that appear in the scene in different quantities. The types are cashew, macadamia, almond, peanut, and pistachio.

For purposes of training, we have manually labelled foreground regions according to nut-type. In Figure 11(c) we have plotted 100 randomly selected pixel-spectra for each type of nut. From the plots, it is not obvious how the spectra of different nuts can be employed for recognition purposes. Indeed, this is a very challenging classification task due to the high degree of similarity between the spectra of different types of nuts. Moreover, the intensity information of the nuts spectra is not reliable due to the variation of the illuminant across the imagery. Further, photometric distortion is also present. This is caused by the viewing geometries for the image pair, which has been acquired using two different sets of cameras, i.e. one for the VIS view and the other one for the NIR image. This leads to corruption due to mismatches in the spectral values during the concatenation of pixel spectra of corresponding points in the two images. This can be clearly appreciated on some of the sample spectra plotted in Figure 11(c), where a “gap” at the 650nm band is visible in a number of traces. This may influence the performance

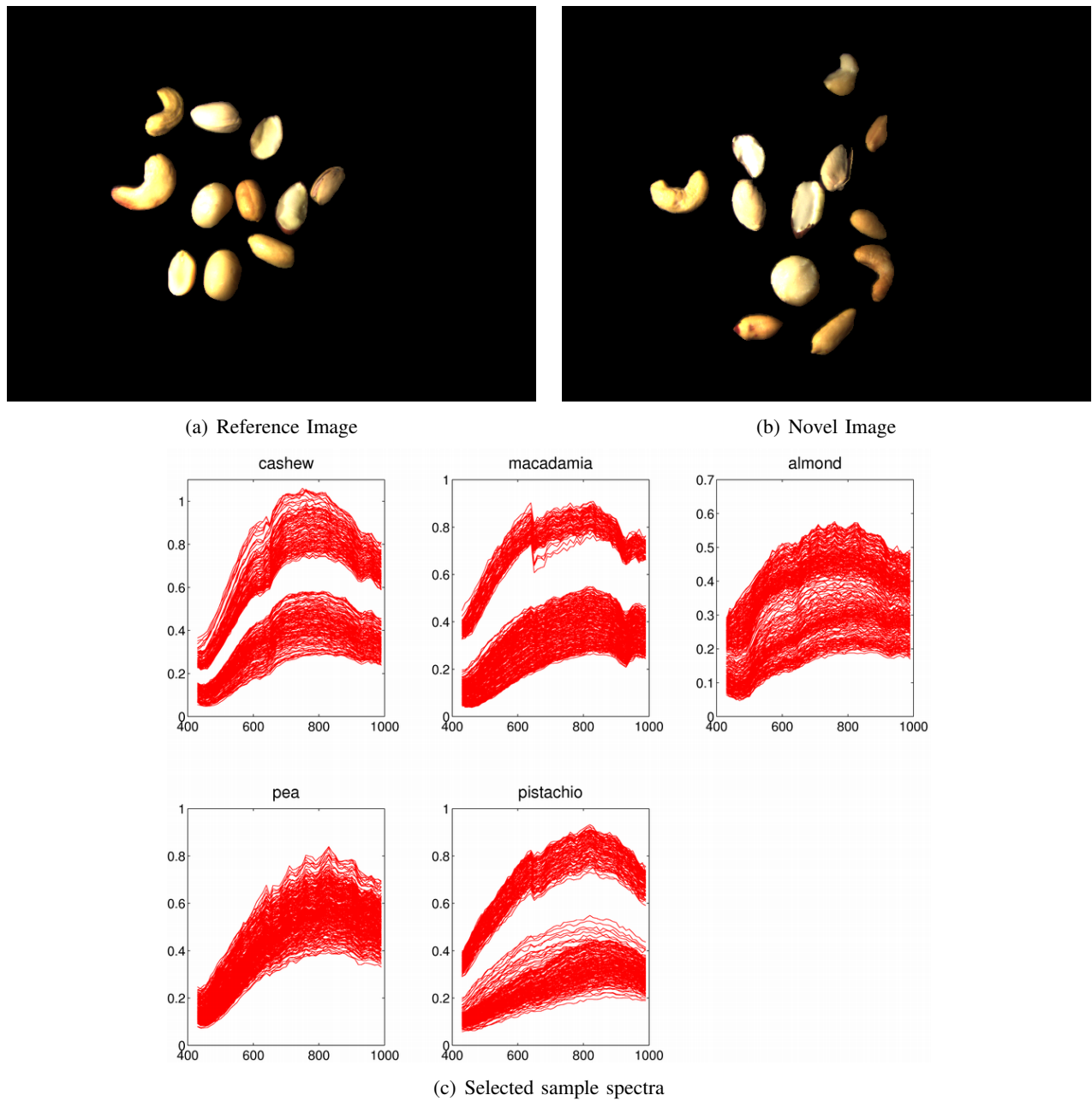


Fig. 11: Pseudo-colour images of assorted nuts illuminated from different directions and corresponding sample spectra.

of SVM classification for raw spectra. Our approach, on the other hand, is based on localised analysis of absorption bands, which renders our method less sensitive to registration artifacts.

Figure 12 shows example surface maps recovered by our method and the alternative for the two images in Figure 11. In the figure, we have used a random training sample with 50 pixels for each nut-type. The left-hand column shows the results on the reference image, while the right-hand column presents the results on the novel image. The ground truth maps, our results and the results yielded by the alternative are displayed in the top, middle, and bottom panels, respectively. Different types of nuts are shown in different colours. From the figure, it is clear that

our algorithm recovers a more plausible material map, i.e. one which is a closer match to the ground truth, than that delivered by the alternative. This is particularly evident in the results on the novel image.

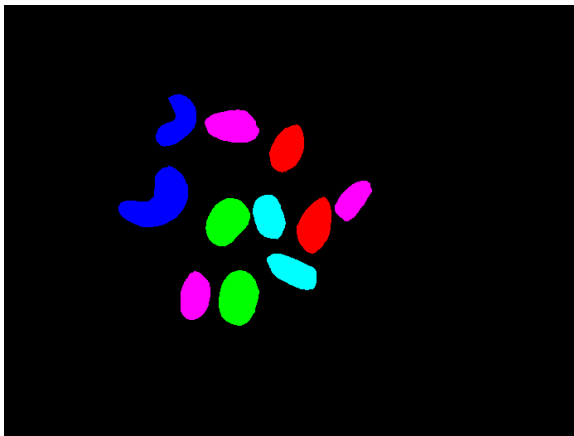
For purposes of a quantitative analysis, here, we adopt the same experimental protocol as the experiment on the wheat data presented earlier. To provide a quantitative analysis, we have randomly selected pixels from different nuts in Figure 11(a). We have then trained the SVM classifiers using both, the absorption features delivered by our algorithm and the raw spectra. For our analysis, we have performed training-test operations with a number of training-set sizes for each class, ranging from 10 to 200 samples per nut-type.

TABLE III: Average accuracy rates and corresponding standard deviations on nut classification for our approach (ABS+SVM) and the alternative based upon raw spectra (SVM).

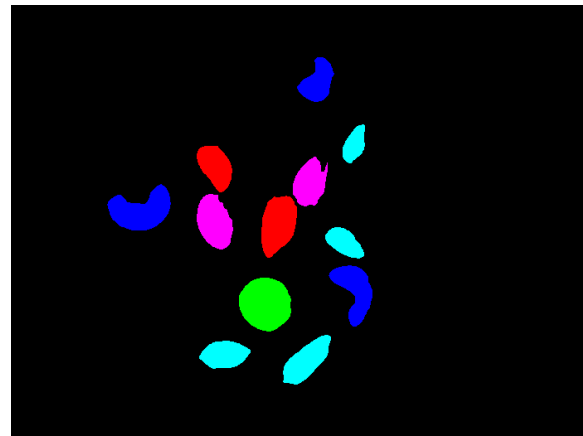
Sample size	Error rate(%)			
	Reference Image		Novel Image	
	SVM	<b>ABS+SVM</b>	SVM	<b>ABS+SVM</b>
10	68.44 $\pm$ 2.56	82.62 $\pm$ 1.60	63.89 $\pm$ 2.38	73.63 $\pm$ 1.82
20	76.32 $\pm$ 1.63	86.59 $\pm$ 1.46	70.69 $\pm$ 2.14	79.14 $\pm$ 1.71
50	80.24 $\pm$ 1.12	87.99 $\pm$ 1.17	76.92 $\pm$ 1.30	81.86 $\pm$ 1.41
100	80.93 $\pm$ 1.04	89.06 $\pm$ 0.97	78.78 $\pm$ 1.11	84.09 $\pm$ 1.49

Table III outlines the classification accuracy rates and standard deviations for our algorithm (ABS+SVM) and the alternative (SVM) with different training sample-set size. From Table III, we can see that our method consistently achieves a better and more stable performance as compared to the direct application of SVMs to the raw spectra, regardless of the size of the training data set. This applies to both the reference image from which training samples were taken and the novel testing image acquired under different photometric condition. For our discriminant absorption-based method, the features extracted are intrinsic to the photometric properties of the material under study and, thus, intra-class variation and noise corruption are greatly reduced.

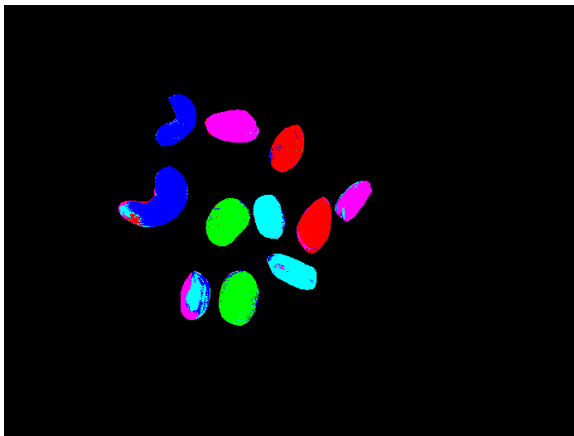
We now conclude the section by examining the features used for classification. Since our algorithm selects features of different lengths for each pairwise SVM classifier, we have applied Kernel principal component analysis (KPCA) [6] to the kernel matrix induced by the feature vectors and kept the first two principal components for visualisation purposes. The distributions of the KPCA projections for the absorption features extracted for each pair of classes are plotted in the left-hand column of Figure 13. In the figure, we have used a random training set of 100 spectra for each nut-type. For purposes of comparison, we also plot the KPCA for the raw spectra on the left-hand column of the figure. For the sake of brevity, we only show the feature projection maps for nut-type pairs corresponding to cashew against macadamia, almond, peanut and pistachio. In all the panels, the blue dots represent feature projections for the cashews, whereas the green dots correspond to the other nut-types. From the feature projections, we can appreciate that the invariant absorption features have better discriminability capacity than the raw spectra. This is exemplified by the large overlap between feature projections from different labels in the KPCA map of raw features, where classes, i.e. nut-type pairs, are scattered. In contrast, the plots corresponding



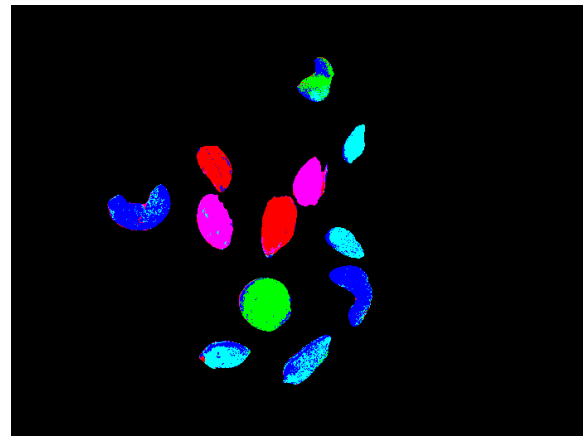
(a) Ground truth (reference image)



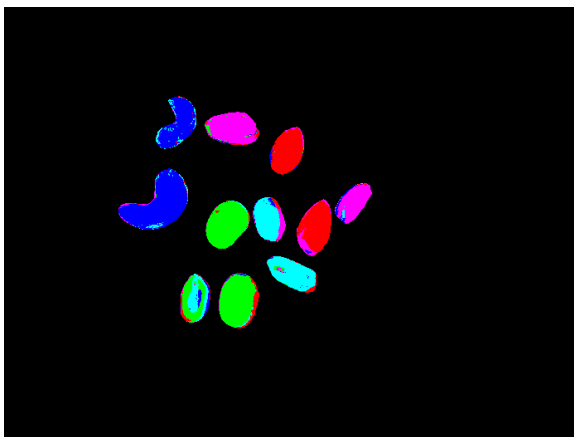
(b) Ground truth (novel image)



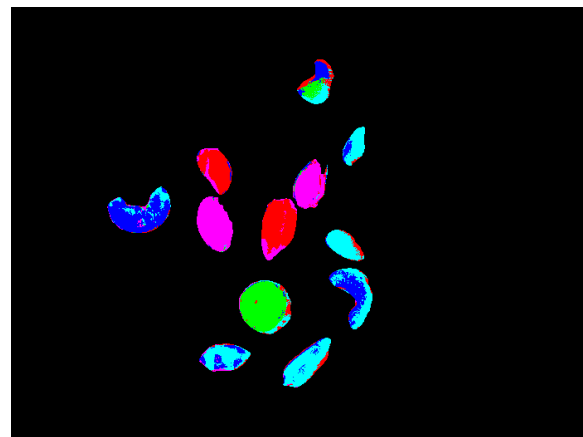
(c) ABS+SVM (reference image)



(d) ABS+SVM (novel image)



(e) SVM (reference image)



(f) SVM (novel image)

Fig. 12: Mapping results for nut images. Top row: Ground truth (cashew: blue, macadamia: green, almond: red, peanut: cyan, pistachio: magenta); Middle row: Surface maps delivered by our algorithm; Bottom row: Surface maps yielded by the alternative.

to our method show better linearly separable mappings, which can then be employed by a large margin classifier like the SVM for improved classification performance.

## VII. CONCLUSION

In this paper, we have presented a novel recognition method for invariant material identification in hyperspectral imagery for both terrestrial and remote sensing settings. To this end, we have adopted an absorption based feature representation and developed a method in which the optimal absorption band-segments are automatically recovered from training data via statistical analysis. Thus, by making use of these invariant absorption features, we have characterised the dissimilarity between materials via probabilistic distributions in the setting of supervised learning. This is achieved by employing the Rényi Divergence, where dissimilarity measures are modelled through the separability between different material classes with optimal parameters recovered by a two step optimisation process. The main contributions of the paper are in the novel photometrically invariant representation of image spectra and a method for feature selection which is based upon a measure of probabilistic dissimilarity. It is worth stressing that the invariance of the absorption features presented here opens-up the possibility of recognising object materials in an ample set of illumination settings. Also, the feature extraction method presented here is quite general in nature and is not limited to absorption features. We have illustrated the utility of the method for purposes of identification on real world imagery, where our algorithm compares favorably to a baseline SVM classifier operating on raw spectral data.

A few directions can be taken for future study. As mentioned in Section III, we have not explicitly taken into account shadow artifacts in the invariant feature representation. Moreover, we mainly focus our attention on photometric invariance in the terrestrial setting. The atmospheric effects in remote sensing have not been addressed explicitly. Although experimental results show the good performance of the proposed algorithm for the classification of both terrestrial and remotely sensed hyperspectral imagery, it would still be interesting to explicitly consider more physical constraints in the modelling and feature representation presented here. Nevertheless, the proposed discriminant feature selection method is quite general in nature and not limited to any specific feature types. The use of the feature selection method for combining multiple cues could also be an interesting future research topic. Finally, we have focused on classification problems for hyperspectral imagery. It is worth exploring the avenues in which the proposed approach can be naturally extended to supervised spectral unmixing making use of the probabilistic formulation of the SVM classifier [51].

## VIII. ACKNOWLEDGMENT

The authors are indebted with the CRC for National Plant Biosecurity, Australia, for facilitating them the wheat samples used in the experimental section of this paper. The work presented here was done while Z. Fu was with the Research School of Information Sciences and Engineering at the Australian National University.

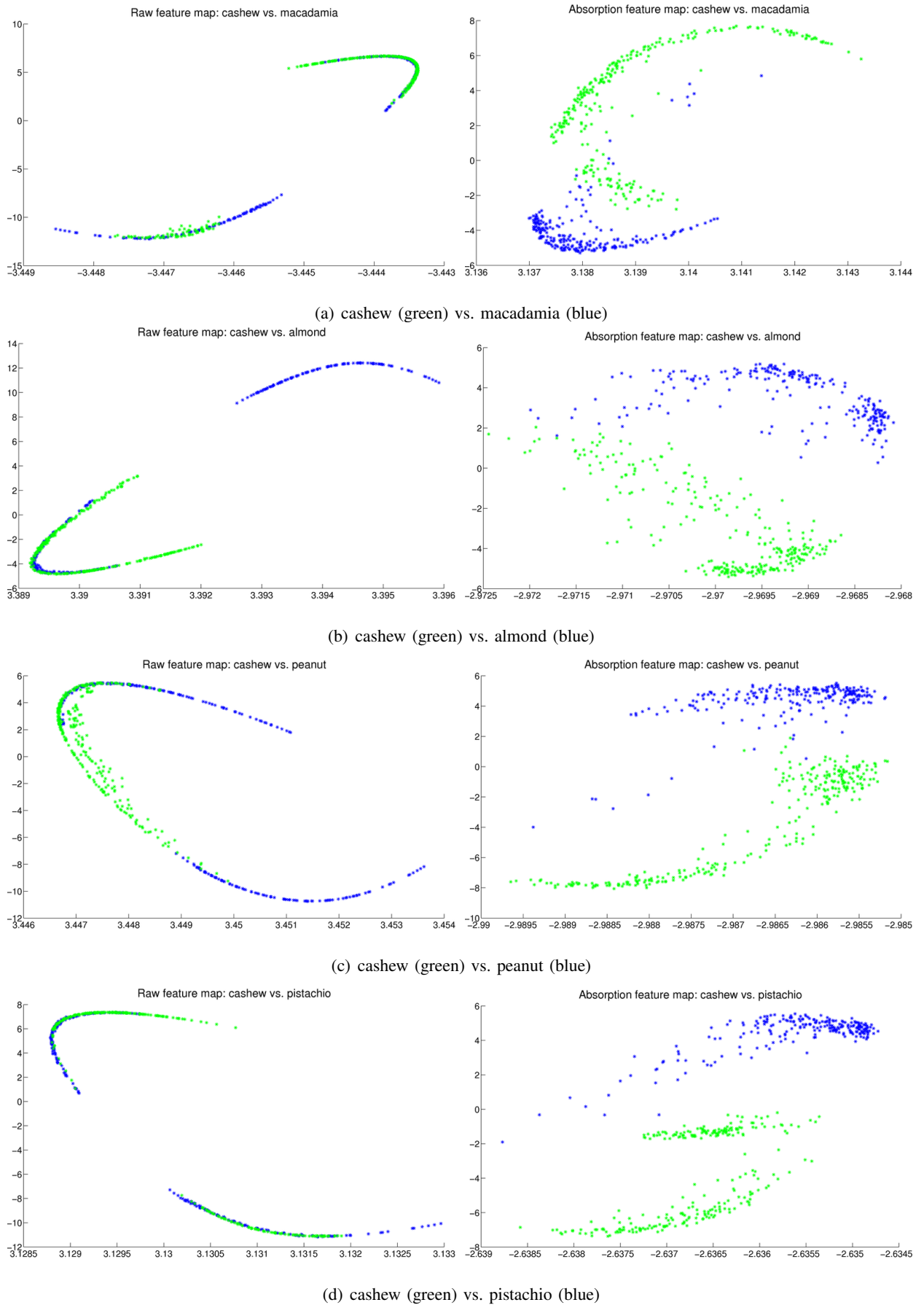


Fig. 13: KPCA mappings for sample nut-type pairs used in our experiments. Left-hand column: KPCA embeddings for the raw spectra; Right-hand column: KPCA projections for our invariant absorption features.

## REFERENCES

- [1] J. Y. Chang, K. M. Lee, and S. U. Lee, "Shape from shading using graph cuts," in *Proc. of the Int. Conf. on Image Processing*, 2003.
- [2] D. Landgrebe, "Hyperspectral image data analysis," *IEEE Signal Processing Magazine*, vol. 19, pp. 17–28, 2002.
- [3] I. Jolliffe, *Principal Component Analysis*. Springer, 2002.
- [4] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2<sup>nd</sup> ed. Academic Press, NY, 1990.
- [5] L. Jimenez and D. Landgrebe, "Hyperspectral data analysis and feature reduction via projection pursuit," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, no. 6, pp. 2653–2667, 1999.
- [6] B. Schölkopf, A. J. Smola, and K.-R. Müller, "Kernel principal component analysis," *Advances in kernel methods: support vector learning*, pp. 327–352, 1999.
- [7] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. Muller, "Fisher discriminant analysis with kernels," in *In IEEE Neural Networks for Signal Processing Workshop*, 1999, pp. 41–48.
- [8] M. Dundar and D. Landgrebe, "Toward an optimal supervised classifier for the analysis of hyperspectral data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 1, pp. 271–277, 2004.
- [9] B. Scholkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2001.
- [10] J. Sunshine, C. M. Pieters, and S. F. Pratt, "Deconvolution of mineral absorption bands: An improved approach," *Journal of Geophysical Research*, vol. 95, no. B5, pp. 6955–6966, 1990.
- [11] B. Hapke, *Theory of Reflectance and Emittance Spectroscopy Topics in Remote Sensing*. Cambridge University Press, 1993.
- [12] R. Clark, G. Swayze, K. Livo, R. Kokaly, S. Sutley, J. Dalton, R. McDougal, and C. Gent, "Imaging spectroscopy: Earth and planetary remote sensing with the usgs tetracorder and expert system," *Journal of Geophysical Research*, vol. 108, no. 5, pp. 1–44, 2003.
- [13] R. N. Clark, A. J. Gallagher, and G. Swayze, "Material absorption band depth mapping of imaging spectrometer data using a complete band shape least-squares fit with library reference spectra," in *Proceedings of the Second Airborne Visible/Infrared Imaging Spectrometer Workshop*, 1990, pp. 176–186.
- [14] J. B. Adams and M. O. Smith, "Spectral mixture modelling: a new analysis of rock and soil types on the viking lander 1 site," *Journal of Geophysical Research*, vol. 91, pp. 8098–8112, 1986.
- [15] J. Harsanyi and C.-I. Chang, "Hyperspectral image classification and dimensionality reduction: An orthogonal subspace projection," *IEEE Trans. on Geosciences and Remote Sensing*, vol. 32, no. 4, pp. 779–785, 1994.
- [16] T. Nichols, J. Thomas, W. Kober, and V. Velten, "Interference-invariant target detection in hyperspectral images," in *Proc. of SPIE*, vol. 3372, 1998, pp. 176–187.
- [17] B. E. Boser, I. Guyon, and V. Vapnik, "A training algorithm for optimal margin classifiers," in *ACM Conf. Computational Learning Theory*, 1992, pp. 144–152.
- [18] J. A. Gualtieri and R. F. Crompt, "Support vector machines for hyperspectral remote sensing classification," in *SPIE 3584*, 1998, pp. 221–232.
- [19] C. Shah, P. Watanachaturaporn, M. Arora, and P. Varshney, "Some recent results on hyperspectral image classification," in *IEEE Workshop on Advances in Techniques for Analysis of Remotely Sensed Data*, 2003.
- [20] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 42, no. 8, pp. 1778–1790, 2004.
- [21] G. Camps-Valls and L. Bruzzone, "Kernel-based methods for hyperspectral image classification," *IEEE Trans. Geoscience and Remote Sensing*, vol. 43, no. 6, pp. 1351–1362, 2005.
- [22] G. Camps-Valls, L. Gomez-Chova, J. Munoz-Mari, J. Vila-Frances, and J. Calpe-Maravilla, "Composite kernels for hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 3, pp. 93–97, 2006.
- [23] L. Bruzzone, M. Chi, and M. Marconcini, "A novel transductive svm for the semisupervised classification of remote sensing images," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 44, no. 11, pp. 3363–3373, 2006.

- [24] Z. Fu, T. Caelli, N. Liu, and A. Robles-Kelly, "Boosted band ratio feature selection for hyperspectral image classification," in *Proc. Intl. Conf. Pattern Recognition*, vol. 1, 2006, pp. 1059–1062.
- [25] G. Healey and D. Slater, "Invariant recognition in hyperspectral images," in *Proc. IEEE Computer Vision and Pattern Recognition*, vol. 1, 1999, pp. 438–443.
- [26] D. Slater and G. Healey, "Material classification for 3d objects in aerial hyperspectral images," in *Proc. IEEE Computer Vision and Pattern Recognition*, vol. 2, 1999, pp. 268–273.
- [27] P. H. Suen, G. Healey, and D. Slater, "The impact of viewing geometry on vision through the atmosphere," in *Proc. IEEE Int. Conf. on Computer Vision*, vol. 2, 2001, pp. 454–459.
- [28] H. Stokman and T. Gevers, "Detection and classification of hyper-spectral edges," in *British Machine Vision Conference*, 1999.
- [29] E. Angelopoulou, "Objective colour from multispectral imaging," in *European Conf. on Computer Vision*, 2000.
- [30] Z. Fu, R. Tan, and T. Caelli, "Specular free spectral imaging using orthogonal subspace projection," in *Proc. Intl. Conf. Pattern Recognition*, vol. 1, 2006, pp. 812–815.
- [31] Chein-I Chang, *Hyperspectral Imaging: Techniques for Spectral Detection and Classification*. Springer, 2006.
- [32] T. Lillesand and W. Ralph, *Remote Sensing and Image Interpretation*, 4<sup>th</sup> ed. John Wiley and Sons, 2000.
- [33] M. Piech and K. Piech, "Symbolic representation of hyperspectral data," *Applied Optics*, vol. 26, pp. 4018–4026, 1987.
- [34] P. Hsu, "Spectral feature extraction of hyperspectral images using wavelet transform," PhD thesis, Dept. of Survey Engineering, National Cheng Kung University, 2003.
- [35] Z. Fu, A. Robles-Kelly, T. Caelli, and R. T. Tan, "On automatic absorption detection for imaging spectroscopy: A comparative study," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 45, no. 11, pp. 3827–3844, 2007.
- [36] S. Shafer, "Using color to separate reflection components," *Color Research and Applications*, vol. 10, pp. 210–218, 1985.
- [37] J. J. Koenderink and A. J. van Doorn, "Geometrical modes as a general method to treat diffuse interreflections in radiometry," *Journal of the Optical Society of America*, vol. 73, no. 6, pp. 843–850, 1983.
- [38] J. A. Marchant and C. M. Onyango, "Shadow-invariant classification for scenes illuminated by daylight," *Journal of the Optical Society of America*, vol. 17, no. 11, pp. 1952–1961, 2000.
- [39] D. Tanre, M. Herman, P. Y. Deschamps, and A. d. Leffe, "Atmospheric modeling for space measurements of ground reflectances, including bidirectional properties," *Applied Optics*, vol. 18, pp. 3587–3594, 1979.
- [40] B. Hapke, "Bidirectional reflectance spectroscopy. iii - correction for macroscopic roughness icarus," *Icarus*, vol. 59, pp. 41–59, 1984.
- [41] F. Schmidt, S. Doute, and B. Schmitt, "Wavanglet: An efficient supervised classifier for hyperspectral images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 5, pp. 1374–1385, 2007.
- [42] B.-C. Gao, K. B. Heidebrecht, and A. F. H. Goetz, "Derivation of scaled surface reflectance from aviris data," *Remote Sensing of Environment*, vol. 44, pp. 165–178, 1993.
- [43] Z. Qu, B. C. Kindel, and A. F. H. Goetz, "The high accuracy atmospheric correction for hyperspectral data (hatch) model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 41, no. 9, pp. 1223–1231, 2003.
- [44] A. Renyi, "On measures of information and entropy," in *4th Berkeley Symposium on Mathematics, Statistics and Probability*, 1960, pp. 547–561.
- [45] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bulletin of the Calcutta Mathematical Society*, vol. 35, pp. 99–109, 1943.
- [46] M. S. Alencar and F. M. Assis, "A relation between the renyi distance of order  $\alpha$  and the variational distance," in *Int. Telecommunications Symposium*, 1998, pp. I:242 – 244.
- [47] H. Jeffreys, "An invariant form for the prior probability in estimation problems," *Proceedings of the Royal Society A*, no. 186, pp. 453–461, 1946.
- [48] R. S. Varga, *Matrix Iterative Analysis*, 2nd ed. Springer, 2000.
- [49] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.
- [50] T. Hastie and R. Tibshirani, "Classification by pairwise coupling," in *Proc. of Neural Information Processing Systems*, 1997.



- [51] J. Platt, "Probabilities for support vector machines," in *Advances in Large Margin Classifiers*, 1999.
- [52] N. Gat, "Imaging spectroscopy using tunable filters: A review," in *SPIE Conf. Algorithms for Multispectral and Hyperspectral Imagery VI*, vol. 4056, 2000, pp. 50–64.
- [53] C. Chang and C. Lin, "Libsvm: a library for support vector machine," 2001, <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [54] C. Hernández Esteban and F. Schmitt, "Silhouette and stereo fusion for 3d object modeling," *Computer Vision and Image Understanding*, vol. 96, no. 3, pp. 367–392, 2004.