Section II

Hyperspectral Image Classification Methods and Approaches

3 Advances in Hyperspectral Image Classification Methods for Vegetation and Agricultural Cropland Studies

Edoardo Pasolli, Saurabh Prasad, Melba M. Crawford, and James C. Tilton

CONTENTS

3.1 INTRODUCTION

Hyperspectral data are becoming more widely available via sensors on airborne and unmanned aerial vehicle (UAV) platforms, as well as proximal platforms. While space-based hyperspectral data continue to be limited in availability, multiple spaceborne Earth-observing missions on traditional platforms are scheduled for launch, and companies are experimenting with small satellites for constellations to observe the Earth, as well as for planetary missions. Land cover mapping via classification is one of the most important applications of hyperspectral remote sensing and will increase in significance as time series of imagery are more readily available.

However, while the narrow bands of hyperspectral data provide new opportunities for chemistry-based modeling and mapping, challenges remain. Hyperspectral data are high dimensional, and many bands are highly correlated or irrelevant for a given classification problem. For supervised classification methods, the quantity of training data is typically limited relative to the dimension of the input space. The resulting Hughes phenomenon [1], often referred to as the curse of dimensionality, increases potential for unstable parameter estimates, overfitting, and poor generalization of classifiers [2]. This is particularly problematic for parametric approaches such as Gaussian maximum likelihood–based classifiers that have been the backbone of pixel-based multispectral classification methods. This issue has motivated investigation of alternatives, including regularization of the class covariance matrices [3], ensembles of weak classifiers [4,5], development of feature selection and extraction methods [6], adoption of nonparametric classifiers, and exploration of methods to exploit unlabeled samples via semi-supervised [7] and active learning [8,9]. Data sets are also quite large, motivating computationally efficient algorithms and implementations.

This chapter provides an overview of the recent advances in classification methods for mapping vegetation using hyperspectral data. Three data sets that are used in the hyperspectral classification literature (e.g., Botswana Hyperion satellite data and AVIRIS airborne data over both Kennedy Space Center and Indian Pines) are described in Section 3.2 and used to illustrate methods described in the chapter. An additional high-resolution hyperspectral data set acquired by a SpecTIR sensor on an airborne platform over the Indian Pines area is included to exemplify the use of new deep learning approaches, and a multiplatform example of airborne hyperspectral data is provided to demonstrate transfer learning in hyperspectral image classification.

Classical approaches for supervised and unsupervised feature selection and extraction are reviewed in Section 3.3. In particular, nonlinearities exhibited in hyperspectral imagery have motivated development of nonlinear feature extraction methods in manifold learning, which are outlined in Section 3.3.1.4. Spatial context is also important in classification of both natural vegetation with complex textural patterns and large agricultural fields with significant local variability within fields. Approaches to exploit spatial features at both the pixel level (e.g., co-occurrence–based texture and extended morphological attribute profiles [EMAPs]) and integration of segmentation approaches (e.g., HSeg) are discussed in this context in Section 3.3.2.

Recently, classification methods that leverage nonparametric methods originating in the machine learning community have grown in popularity. An overview of both widely used and newly emerging approaches, including support vector machines (SVMs), Gaussian mixture models, and deep learning based on convolutional neural networks is provided in Section 3.4. Strategies to exploit unlabeled samples, including active learning and metric learning, which combine feature extraction and augmentation of the pool of training samples in an active learning framework, are outlined in Section 3.5. Integration of image segmentation with classification to accommodate spatial coherence typically observed in vegetation is also explored, including as an integrated active learning system. Exploitation of multisensor strategies for augmenting the pool of training samples is investigated via a transfer learning framework in Section 3.5.1.2. Finally, we look to the future, considering opportunities soon to be provided by new paradigms, as hyperspectral sensing is becoming common at multiple scales from ground-based and airborne autonomous vehicles to manned aircraft and space-based platforms.

3.2 HYPERSPECTRAL TESTBED FOR VEGETATION CLASSIFICATION

Five publically available hyperspectral benchmark data sets, which have been used to evaluate classification algorithms for vegetation-based studies in the literature, are included to illustrate the methodology presented in this chapter. The data were acquired by space-based and airborne sensors covering the visible and short-wave infrared portions of the spectrum at spatial resolutions ranging from 2 to 30 m. The higher spatial resolution acquisition allows both discrimination of smaller objects and utilization of texture information by the classification algorithms. Spectral signatures of the classes are complex and often overlapping, and spatial patterns include agricultural fields with regular boundaries and natural vegetation where classes are often fragmented or mixed. Characteristics of the data sets are listed in Table 3.1.

3.2.1 Botswana Hyperion Data (BOT)

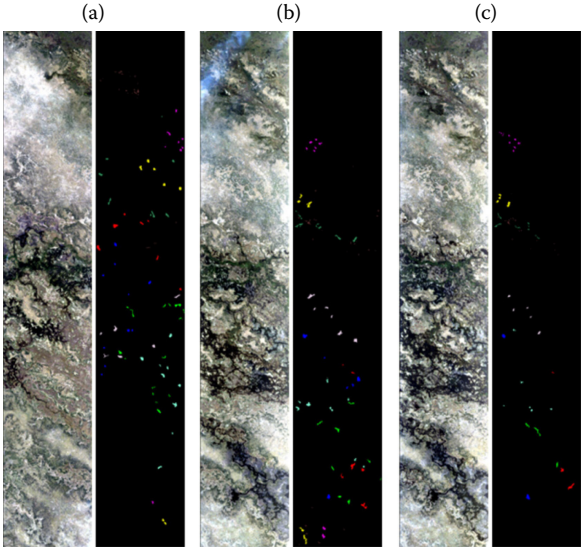
The NASA EO-1 satellite was launched in November 2000, with Hyperion as an "auxiliary" hyperspectral sensor. The EO-1 platform was designed for one year of operation, but was finally decommissioned nearly two decades later in 2017. Hyperion data were acquired in 7.7-km strips at 30 m spatial resolution for a multiyear study of flooding in the Okavango Delta, Botswana. Uncalibrated and noisy bands were removed, leaving 145 bands as candidate features for classification. Nine classes of complex natural vegetation were identified by researchers at the Okavango Research Center. Class groupings include seasonal swamps, occasional swamps, and woodlands, and are distributed in fragmented patterns over a large area. RGB images of the area, maps of the ground reference data, and a class legend are included in Figure 3.1. As shown in the figure, the pointing angle of the satellite changed after the May acquisition, necessitating development of knowledge transfer models. Signatures of several classes overlap spectrally, resulting in a challenging data set for classification. Class 3 (riparian) and Class 6 (woodlands) are particularly difficult to discriminate. After removing water absorption and noisy and overlapping spectral bands in the visible and near-infrared (VNIR) and short-wave infrared (SWIR) sensors, 145 bands of May and July 2001 images were used for classification experiments.

3.2.2 Kennedy Space Center AVIRIS Data (KSC)

Airborne hyperspectral data were acquired by NASA AVIRIS at 18 m spatial resolution and 10 nm spectral resolution over a natural wetland/upland environment adjacent to the Kennedy Space Center, Florida, in March 1996, to evaluate the impact of drainage management practices on the incursion of invasive species into an endangered species habitat. Figure 3.2 includes an RGB image of the area, ground reference data, and class reference information. Noisy and water absorption bands

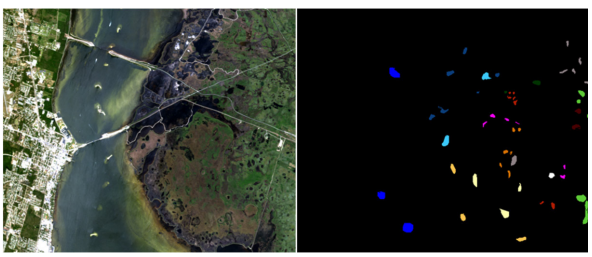
TABLE 3.1
Testbed Data Sets and Associated Characteristics

Data Set	BOT	KSC	Indian Pine 1992	Indian Pine 2010	Galveston, Texas	Galveston, Texas
Scene description	Vegetation and flooding	Wetland/upland vegetation	Early season agriculture	Early season agriculture	Wetland vegetation	Wetland vegetation
Sensor	Hyperion	AVIRIS	AVIRIS	SpecTIR	SpecTIR	Headwall nano
Platform	Satellite	Airborne	Airborne	Airborne	Airborne	Terrestrial
Spectral range	0.4–2.5 μm	0.4–2.5 μm	0.4–2.5 μm	0.4–2.5 μm	0.4–2.5 μm	0.4–1.0 μm
Spatial resolution	30 m	18 m	18 m	2 m	1 m	Variable
No. of bands	220	176	176	360	360	274
No. of classes	9	13	16	12	12	12



ID	Class	#Labeled Samples May	June	July
	C1: Water	297	361	185
	C2: Primary floodplain	437	308	96
	C3: Riparian	448	303	164
	C4: Firescar	354	335	186
	C5: Island interior	337	370	131
	C6: Woodlands	357	324	169
	C7: Savanna	330	342	171
	C8: Short mopane	239	299	152
	C9: Exposed soils	215	229	96

FIGURE 3.1 Botswana (BOT) Hyperion data. True color composites with corresponding ground reference data in (a) May, (c) June, and (e) July 2001, class labels and # labeled samples. The pointing angle changed after the May acquisition, and then remained the same for subsequent dates.



ID	Class	#Labeled Samples
	C1: Scrub	761
	C2: Willow swamp	243
	C3: Cabbage palm hammock	256
	C4: Cabbage palm/Oak	252
	C5: Slash pine	161
	C6: Oak/Broadleaf hammock	229
	C7: Hardwood swamp	105
	C8: Graminoid marsh	431
	C9: Spartina marsh	520
	C10: Cattail marsh	404
	C11: Salt marsh	419
	C12: Mud flats	503
	C13: Water	927

FIGURE 3.2 Kennedy Space Center (KSC) AVIRIS data. RGB true-color composite and corresponding ground reference map, class labels, and # labeled samples.

were removed from the reflectance data, leaving 176 features for 13 wetland and upland classes. The spectral signatures of multiple classes are mixed and often exhibit only subtle differences. Cabbage Palm Hammock (Class 3) and Broad Leaf/Oak Hammock (Class 6) are upland trees; Willow Swamp (Class 2), Hardwood Swamp (Class 7), Graminoid Marsh (Class 8), and Spartina Marsh (Class 9) are trees and grasses in transition wetlands. Classification results for all 13 classes and for these difficult classes are reported in several publications.

3.2.3 Indian Pine AVIRIS 1992 and SpecTIR 2010 Data

The historical data acquired by the NASA AVIRIS sensor in June 1992 over a central Indiana farming area have been widely used to evaluate classification methods that exploit spatial information. After removing 20 water absorption bands, 200 bands are used for analysis. The scene is composed of agricultural fields with regular geometry, providing an opportunity to evaluate the impact of within-class and between-class variability at medium spatial resolution. The spectral signatures of corn and soybean fields, which were planted only a short time prior to the acquisition, illustrate the impact of tillage management practices.

The 16 classes of labeled reference data are reported at the field scale for crops, but significant within-field variability resulted in heterogeneous spectral responses. Although labeled as vegetation, the spectral responses of many classes are dominated by soil and residue signatures from the previous year. Figure 3.3 includes an RGB image of the area, class legends, and the corresponding labeled data.

Additional hyperspectral imagery was acquired by the airborne ProSpecTIR VS2 VNIR/SWIR in June 2010 for a study of residue cover estimates over an area near the location of the original Indian Pine AVIRIS data. The data were collected at 2 m spatial resolution in 360 channels at 5 nm spectral resolution over the range of 390–2450 nm. Bands were composited to 10 nm, and 178 spectral bands were used for analysis. Nineteen classes of crops, residue, and buildings were identified. A true color composite is shown in Figure 3.4 with associated ground reference data and class information.

3.2.4 Texas Coast SpecTIR Data

A heterogeneous hyperspectral data set composed of airborne imagery and a unit collected over wetland vegetation are included to illustrate domain adaptation (transfer learning) (Figure 3.5).

ID	Class	# Labeled Samples
	C1: Corn-no till	1434
	C2: Corn-min till	834
	C3: Corn	234
	C4: Soybeans-no till	968
	C5: Soybeans-min till	2468
	C6: Soybeans-clean till	614
	C7: Alfalfa	54
	C8: Grass/Pasture	497
	C9: Grass/Trees	747
	C10: Grass/Pasture-mowed	26
	C11: Hay-windrowed	489
	C12: Oats	20
	C13: Wheat	212
	C14: Woods	1294
	C15: Building/Grass/Trees/Drive	380
	C16: Stone/Steel towers	95

FIGURE 3.3 Indian Pine 1992 AVIRIS data. RGB true-color composite and corresponding ground reference map, class labels, and # labeled samples.

ID	Class	# Labeled Samples
	C1: Corn-high residue	47,168
	C2: Corn-mid residue	183,530
	C3: Corn-low residue	356
	C4: Soybean-high residue	226,130
	C5: Soybean-mid residue	120,400
	C6: Soybean-low residue	29,210
	C7: Other residues	5,795
	C8: Wheat	3,387
	C9: Hay	63,135
	C10: Grass/Pasture	6,512
	C11: Grass	26,853
	C12: Wood-uniform	64,947
	C13: Wood-rugged	288,500
	C14: Highway	10,637
	C15: Local road	6,570
	C16: Power station	6,929
	C17: Power towers	411
	C18: Houses/Buildings	2,128
	C19: Urban areas	1,532

FIGURE 3.4 Indian Pine 2010 SpecTIR data. RGB true-color composite and corresponding ground reference map, class labels, and # labeled samples.

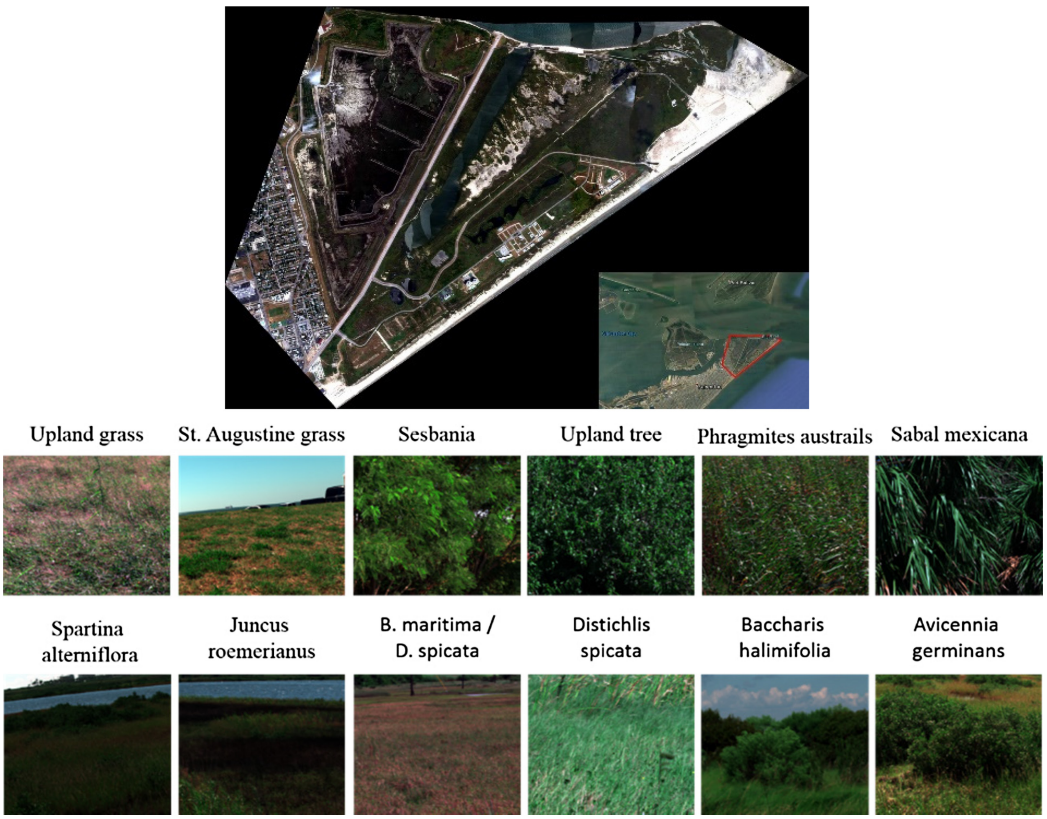


FIGURE 3.5 Galveston, Texas, data. True-color images of the aerial view (target domain) and street view (source domain) wetland data. The location of the study site is indicated by the red box in a Google Earth screenshot image.

Changes in distribution of wetland vegetation species can have profound impacts on the coastal economy and ecology; hence, studying wetland vegetation through remote sensing is of great importance, particularly over extended geographic areas.. Marshes in Misson-Aransas estuary, which were previously dominated by smooth cordgrass (*Spartina alterniflora*), have been replaced by black mangroves (*Avicennia germinans*).

Hyperspectral imagery was acquired by the airborne ProSpecTIR VS sensor with a spatial coverage of 3462 × 5037 pixels at a 1 m spatial resolution. A field survey was undertaken on September 16, 2016. Figure 3.6 depicts the mean spectral signatures of 12 identified species/classes (Table 3.2) As part of this field survey, side-looking hyperspectral imagery (called "street view" imagery in this chapter) was collected using a Headwall Nano-Hyperspec sensor of the same area. It has 274 bands spanning 400–1000 nm at a 3 nm spectral resolution. This resulted in a unique domain adaptation problem where models were trained from very high-resolution street-view (terrestrial) imagery and transferred to aerial imagery. It is also very challenging because the two domains (street view and aerial) are different in many ways, including the viewpoints, illumination conditions, atmospheric conditions, and so on.

3.3 MANAGING THE FEATURE SPACE

3.3.1 Dimensionality Reduction via Feature Selection and Extraction

3.3.1.1 Feature Selection

Identifying features that are effective for modeling data class characteristics is a critical preprocessing step for hyperspectral image classification. Apart from computational

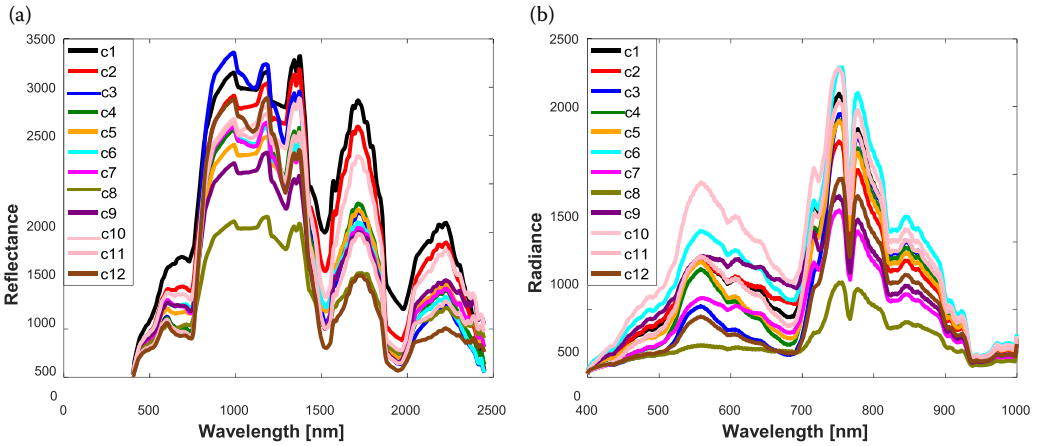


FIGURE 3.6 Galveston, Texas, SpecTIR airborne and Headwall handheld camera data. Mean spectral signature of the (a) aerial view (target domain) and (b) street view (source domain) wetland data.

TABLE 3.2
Galveston, Texas, Data

Class	# Labeled Samples (Aerial View)	# Labeled Samples (Street View)
C1: Upland grass	794	1463
C2: St. Augustine grass	100	1009
C3: Sesbania	294	1021
C4: Upland tree	426	1040
C5: *Phragmites austrails*	780	1029
C6: *Sabal mexicana*	74	1189
C7: *Spartina alterniflora*	733	1152
C8: *Juncus roemerianus*	202	1264
C9: *Batis maritima/Distichlis spicata*	596	1106
C10: *Distichlis spicata*	1197	1087
C11: *Baccharis halimifolia*	360	1017
C12: *Avicennia germinans*	1663	1119

requirements, classifiers tend to have low generalization capabilities when data are characterized by high dimensionality, especially when the number of training samples is limited with respect to the number of features. A traditional solution to deal with this problem is represented by feature selection (FS), which aims to reduce the dimensionality of the original feature space by choosing the best—and ideally the minimum—subset of features. Numerous FS approaches have been proposed in the last few decades [6] and can be grouped in two main categories: *filter* and *wrapper* methods. Filter methods perform FS as a preprocessing step where the selection criterion is independent of the classifiers used to subsequently perform classification of the data [10]. Wrapper strategies perform FS based on the performance of a given classifier [11]. These techniques are generally applied to the original spectral bands, although they can be extended to newly generated [12] or extracted features [13]. Selection from the original feature space is advantageous in the sense that the resulting features retain a physical relationship to the original process. We focus in the following on filter methods, which can be categorized as supervised and unsupervised approaches.

Supervised FS can be further subdivided in two major groups, that is, parametric and nonparametric methods. *Parametric supervised FS* involves class modeling using training data. A widely adopted method is based on the Jeffries-Matusita (J-M) distance, which measures the separability of two class density functions [14]. When classes are assumed to be Gaussian distributed [15], computation of the J-M distance is based on the Bhattacharyya distance. This approach performs well when the Gaussian distribution assumption is valid. It is used to find a subset of features to best accommodate class data variations at multiple sites/locations and generate a visual representation of the separation capability provided by each band, which then leads to quantitative band selection [16]. Other distance measures can also be used, including spectral angle, Euclidean distance, and Mahalanobis distance. Apart from distances between classes, the ratio of within-class and between-class variance can be used as well to define separability [17]. *Nonparametric supervised FS* considers the information provided by the training data directly, without requiring class data modeling. As an example, overlapping and noisy bands can be removed using a canonical correlation measure to obtain an optimal subset of features that provide the best estimate of the center of classes [18]. Information theory can also be applied to perform nonparametric supervised FS. Mutual information (MI) provides a measure of linear and nonlinear dependency between two variables [19] and is suitable for general cases since no assumptions about the shape of the class data density functions must be made.

Unsupervised FS aims at reducing feature redundancy. For example, features that are dissimilar to those already selected can be chosen one by one via linear prediction error analysis [20]. Other methods use similarity measures to partition the original set of features into a number of homogeneous clusters and then select a representative feature from each cluster [21]. MI is used to find the subset of features with minimum dependency [22]. Subsets of representative features can also be selected simultaneously through geometry-based FS methods [23].

After defining the selection criterion, adopting the appropriate *searching strategy* is a challenging task. A first solution is represented by *exhaustive search* in which all the possible combinations are evaluated. Exhaustive search is often impractical, so alternative suboptimal approaches are usually adopted and defined as *greedy search*. These optimization problems are usually not convex, and heuristic strategies are needed. In the case of a monotonic criterion, the branch and bound method [24] can be applied to avoid exhaustive search in a moderate-sized searching space. Sequential forward selection and backward elimination are fast approaches, but do not allow feedback to revise previously selected features. Improvements are represented by sequential forward floating selection and sequential backward floating selection [25] in which the selected features are reconsidered for inclusion or deletion at each iteration. Combinatorial optimization approaches use heuristic methods to reduce the number of features. Proposed solutions include methods based on genetic algorithms [26], particle swarm optimization [27], and clonal selection [28].

3.3.1.2 Spectral Indices

Spectral indices computed as ratios of broad-band spectral bands or of normalized differences between pairs of bands based on multispectral radiance and reflectance data have been used for nearly half a century in studies of vegetation. More recently, narrow-band indices based on spectral bands and derivatives of the reflectance spectrum have been explored in vegetation studies for their value in characterizing biophysical properties, predicting variables of interest, and mapping or unmixing land cover. The continuous spectrum provided by hyperspectral sensors and the relationship of spectral absorption features in spectral signatures to chemistry-based properties provides robust capability to target explicit characteristics compared to what can be achieved by broad-band multispectral indices. Thenkabail et al. [29] conducted a comprehensive review of spectral indices derived from the EO-1 Hyperion hyperspectral sensor using data acquired for a wide range of agricultural applications and geographic sites. Recently, narrow-band indices have also become a focus of high-throughput phenotyping by plant breeders seeking to map and relate phenotypic traits to genotypes [30].

Spectral indices also provide an appealing physically-based capability for reducing the dimensionality and redundancy that are inherent in hyperspectral signatures over vegetation.

Bridging band-specific feature selection and extraction approaches that seek to represent relevant information in the spectrum via global transformations, hyperspectral indices provide robust, local transformations that are useful for a wide range of applications in remote sensing–based studies of vegetation, including classification.

3.3.1.3 Linear Transformation–Based Approaches

Feature extraction (FE) seeks to project the original feature space onto a small number of features. While extracted features lose their relationship to the original physical phenomena, they provide a compressed version of the original feature set. Each feature in the original hyperspectral data is characterized by a contribution determined by the transformation matrix associated with the chosen extraction method. FE techniques can be categorized as unsupervised (global data oriented) and supervised (class data oriented) or as linear and nonlinear.

In *supervised FE* methods, new discriminative features can be obtained by combining the original features into groups. As an example, adjacent correlated features can be combined into a smaller number of features retaining the original spectral interpretation [31]. B-dis can be adopted as a grouping criterion before creating new features as a weighted sum of the features in each group [32]. In [33], contiguous groups of features are averaged based on JM distance. Linear discriminant analysis (LDA) and canonical analysis are traditional parametric FE techniques that are based on the mean vector and covariance matrix of each class. The ratio of within-class to between-class scatter matrices is used to formulate an effective criterion of class separability. Limitations of LDA include its dependence on the distributions of classes being approximately Gaussian and its inability to handle cases where class data do not form a single cluster. Further, the maximum rank of the between-class scatter matrix is the number of classes (M) minus one. Decision boundary FE (DBFE), an early method developed specifically for hyperspectral data [34], aims at finding new features that are normal to class decision boundaries. Nonparametric FE via regularization techniques have also been proposed to overcome the limitations of LDA and obtain more stable results [35]. Nonparametric discriminant analysis (NDA) [36] defines a nonparametric between-class scatter matrix based on a critical finding that samples close to the boundary are more relevant than those far from it. Nonparametric weighted FE (NWFE) was developed in light of NDA, introducing regularization techniques to achieve better performance for hyperspectral image classification than LDA and NDA. Double nearest proportion (DNP) FE builds new scatter matrices based on a double nearest proportion structure [37].

Unsupervised FE is usually obtained by combining the original group of features via an average or weighted sum operation. For example, top-down and bottom-up decompositions can be adopted to merge highly correlated adjacent features and then project them onto their Fisher directions [31]. Apart from combining groups of contiguous features, a more general FE approach consists of mapping the original high-dimensional space into a low-dimensional one via a data transform. Two typical data transformation methods are principal component analysis (PCA), which reduces the original set of features into a smaller set of orthogonal ones computed as linear combinations of the original features with maximum variance [38], and independent component analysis (ICA), a statistical technique for separating the independent signals from overlapping signals [39]. Other techniques seek projections via different optimization models. Projection pursuit (PP) methods search for projections that optimize certain projection indexes [40]. In both supervised and unsupervised FE, the kernel trick is an easy way to extend linear models to nonlinear ones, as we describe in more detail in the next section. In [41–43], angular distance–based supervised, unsupervised, and semisupervised discriminant analysis and their kernel variants were proposed and shown to outperform Euclidean distance–based variants of such algorithms for hyperspectral classification.

3.3.1.4 Manifold Learning

Although hyperspectral data are typically modeled assuming that the data originate from linear processes, and linear feature extraction approaches are simple and straightforward to implement,

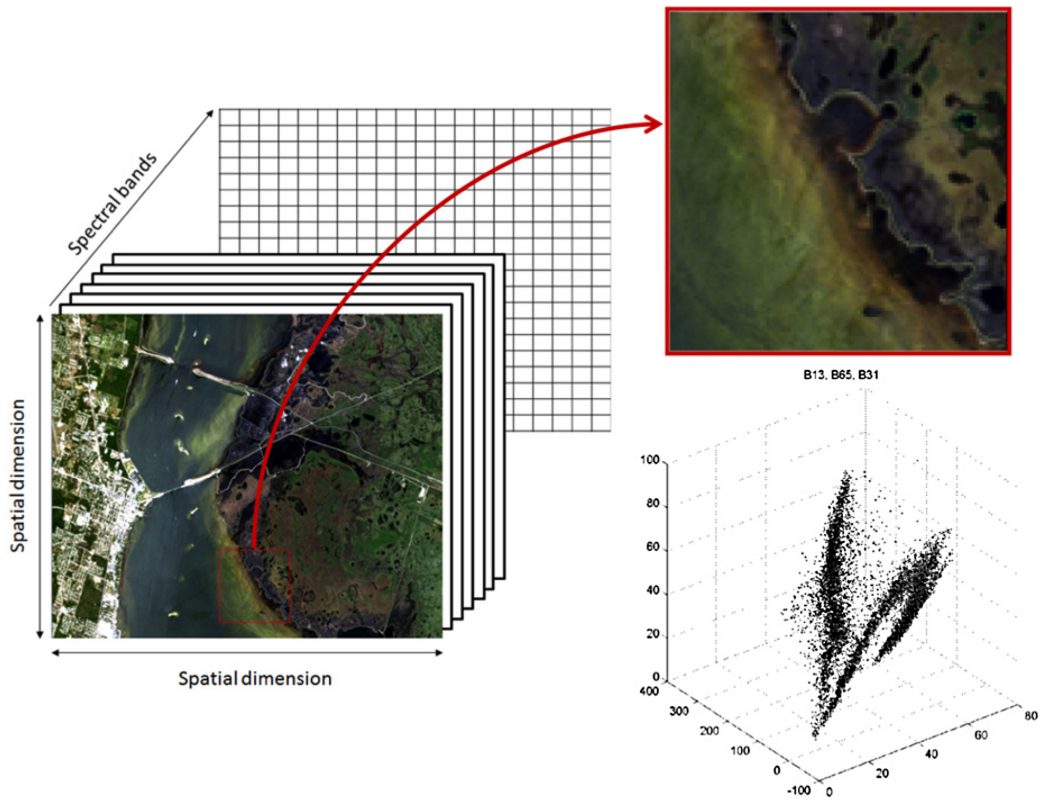


FIGURE 3.7 Nonlinearity in the spectral data is exhibited in a plot of bands 13, 65, and 31 for the Kennedy Space Center data set. (Adapted from L. Bruzzone et al. *Hyperspectral Data Exploitation: Theory and Applications*, pp. 275–311, 2007. [45])

nonlinearities associated with physical processes are often exhibited in the narrow-band data [44]. Nonlinear feature extraction techniques assume that the high-dimensional data inherently lie on a lower-dimensional manifold, as shown in Figure 3.7.

The machine learning community initially demonstrated the potential of manifold-based approaches for nonlinear dimensionality reduction and modeling of nonlinear structures [46–50], and its application to classification of hyperspectral data has been exhibited over the past decade for unsupervised, supervised, and semisupervised learning [45,51], as well as for multitemporal scenarios [52,53]. Nonlinear manifold learning methods are broadly characterized as globally or locally based approaches. Global manifold methods seek to maintain the fidelity of the overall topology of the data set at multiple scales of the data, while local methods preserve local geometry and are computationally efficient because they only require sparse matrix computations. Global manifolds are generally less susceptible to overfitting the data, which is beneficial for generalization in classification, but local methods potentially provide opportunities for better representation of heterogeneous data with submanifolds. Traditional manifold learning methods whose theoretical foundation is associated with the eigenspectrum and kernel framework include isometric feature mapping (ISOMAP) [46], kernel principal component analysis (KPCA) [47], and locally linear embedding (LLE) [48].

Manifold learning is classically represented in a graph-embedding approach, both for efficiency and computational simplicity. A Laplacian regularizer is employed to constrain the classification function to be smooth with respect to the data manifold, and the resulting coordinates are obtained from the eigenvectors of a graph Laplacian matrix. Different manifold learning methods correspond

to specific graph structures. Compared to global methods that represent distances across the full manifold, local manifold learning evaluates the local geometry in each neighborhood, and data points that are included in neighborhood are connected in the graph. In the graph-embedding framework, manifold coordinates are obtained from the eigenvectors of the graph Laplacian matrix. The graph can either be developed from the combined training and testing data (unsupervised), or manifold learning can be applied to the training data and an out-of-sample extension method employed to incorporate the testing data (supervised) [54,55]. The first strategy can provide more accurate manifold coordinates, while the latter is advantageous when the quantity of testing data is large.

In the general formulation of the graph Laplacian-based framework, samples are defined in a data matrix $X = [x_1, x_2, \ldots, x_n]$, $x_i \in \mathbb{R}^m$, where n is the number of samples and m is the feature dimension. The dimensionality reduction problem seeks a set of manifold coordinates $Z = [z_1, z_2 \ldots, z_n]$, $z_i \in \mathbb{R}^p$, where typically, $m >> p$, through feature mapping $\Phi: x \rightarrow y$, which may be analytical (explicit) or data driven (implicit), and linear or nonlinear. Assuming an undirected weighted graph $G = \{X, W\}$ with data samples X and algorithm-dependent similarity matrix W, the graph Laplacian $L = D - W$, with diagonal degree matrix $D_{ii} = \Sigma_j W_{ij}$, $\forall_i$. Given labeled data $X_l = [x_1, x_2, \ldots, x_l]$ and unlabeled data $X_u = [x_{l+1}, x_{l+2}, \ldots, x_{l+u}]$, $\Omega = [\Omega_1, \ldots, \Omega_C]$, where C is the number of classes. Class labels of X_l are denoted as $Y_l \in \mathbb{R}^{C \times l}$ with $Y_{ij} = 1$ if $x_j \in \Omega_i$. The data points $X = [X_l, X_u]$ produce a weighted graph $G = \{X, W\}$, where X consists of $N = l + u$ data points, and $W \in \mathbb{R}^{N \times N}$ is the adjacency matrix of the connected edges between data points. For a directed graph, a symmetric adjacency matrix $W = W + W^T$ can be assumed. The dimensionality reduction criterion can be represented as $Y^* = \arg \min_{YBY^T = I} tr(Y LT^T)$, where B is a constraint matrix that depends on the dimensionality reduction method.

Plots of coordinates obtained using PCA, Isomap, and LLE for the nine-class NASA Hyperion BOT data, with optimal parameters for each manifold embedding, are shown in Figure 3.8.

Differences in the separation of classes, which relate to potential discrimination via classification, are shown for the three projections. The objective functions for neither linear projections such as PCA nor manifold learning are related to the classification objective, so the resulting projections may or may not provide improved separation of classes.

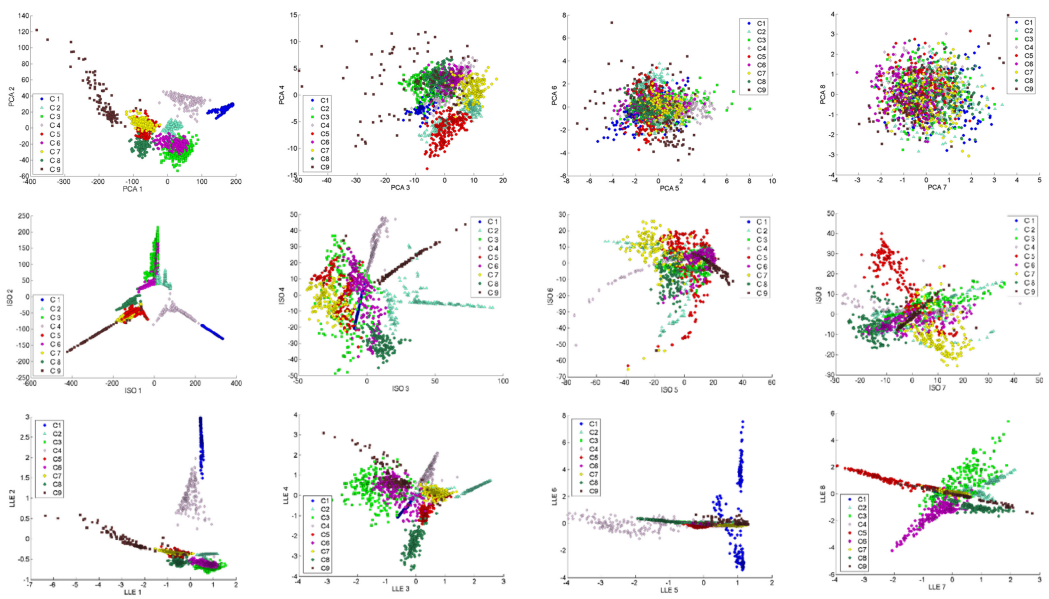


FIGURE 3.8 Plots of coordinates obtained from (first line) PCA, (second line) ISOMAP, and (third line) LLE for the nine Botswana classes (C1-C9). (a) bands 1–2, (b) bands 3–4, (c) bands 5–6, (d) bands 7–8.

3.3.2 Incorporation of Spatial Context

For many applications, a key discriminative component that makes hyperspectral imaging appealing is the richness of the spectral reflectance profiles. There is much to be gained for most applications, including vegetation classification, by leveraging the spatial context in the imagery. A common approach to incorporating spatial context for classification is to extract meaningful morphological and textural features [56,57] and then learn the classifier in the resulting feature space. Within this space, a choice that has been demonstrated to be particularly successful for hyperspectral classification tasks is extended morphological attribute profile (EMAP) features, that represent profiles created by removing connected components that do not meet some criteria specified a priori. For situations where the criteria are satisfied, the regions are kept intact; otherwise, they are set to the gray level of a darker or brighter surrounding region. Attributes (features) can represent the morphology/geometric properties of the objects (e.g., image moments and shape) or the textural information about the objects (e.g., range, standard deviation, and entropy). EMAP features have found significant success in hyperspectral image analysis—in the context of multichannel images, such features are often computed from a subset of features extracted from the hypercube (e.g., selected bands or the first few principal components of the cube). They have been applied to a wide variety of remote sensing applications to extract spatial context for image analysis [58,59].

3.4 IMAGE CLASSIFICATION STRATEGIES

3.4.1 Classical Pixel (Sample)–Based Classification for Hyperspectral Image Analysis

Within the realm of geospatial image analysis in general and hyperspectral image analysis in particular, a significant focus within the research community has been on the design of feature reduction and analysis algorithms (classification, change and anomaly detection, target recognition, spectral unmixing, etc.) [60–63]. Similar to feature extraction and dimension reduction, classifiers that utilize domain-specific properties of hyperspectral data have emerged as popular choices [64–68]. Over the past decade, the choice of classifiers for hyperspectral classification has shifted from traditional approaches such as *K*-nearest neighbors, Gaussian maximum-likelihood classification, and random forests to more advanced approaches that offer greater capacity in modeling nonlinear decision surfaces in the feature space. *K*-nearest neighbor–based classifiers assume that data can be classified by surveying the *K* training data points nearest the test point in the feature space—an assumption that is effective when coupled with local manifold learning approaches, but is far too simplistic as a classification scheme in itself. Likewise, Gaussian maximum-likelihood classifiers assume that class-conditional likelihoods are modeled as Gaussian distributions in the spectral reflectance parameter space, and the underlying parametrization is then learned via maximum likelihood—the approach has several shortcomings, including the necessity to undertake feature reduction as a preprocessing step prior to inferring the high-dimensional parameters and departure of class-conditional distributions from unimodal Gaussian behavior under a variety of practical remote sensing scenarios (such as the same class being present in a well-illuminated part of the scene and under cloud shadows). More advanced classification approaches have hence been proposed and utilized in recent years for hyperspectral classification. These include multikernel learning and variants of support vector machines [69,70], and other model-based classifiers, spectrally constrained implementations of Gaussian mixture models (GMMs) [66], sparse representation classification (SRC) [71,72], local approaches [73], and so on.

3.4.1.1 Single-Kernel Support Vector Machines

Given a training data set $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{n}$ in $\mathbb{R}^d$ with class labels $y_i \in \{+1, -1\}$ and a nonlinear kernel function $\phi(\cdot)$, an SVM [74] classifies data by learning the optimal hyperplane in the Hilbert space induced by the kernel function

$$\min_{\omega,\xi_i,b}\left\{\frac{1}{2}\|\omega\|^2 + \varsigma\sum_{i=1}^{n}\xi_i\right\} \tag{3.1}$$

subject to the constraints

$$y_i(\langle\phi(\boldsymbol{\omega},\mathbf{x}_i)\rangle+b)\geq 1-\xi_i, \tag{3.2}$$

for $\xi_i \geq 0$ and $i = 1, \ldots, n$, where ω is normal to the optimal decision hyperplane [i.e., $\langle\omega, \phi(\mathbf{x})\rangle + b = 0$], n denotes the number of samples, b is the bias term, ς is the regularization parameter that controls the generalization capacity of the machine, and ξ_i is the positive slack variable allowing appropriate accommodation of permitted errors. The above problem is solved by maximizing its Lagrangian dual form,

$$\max_{\alpha}\left\{\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_i y_i y_j K(\mathrm{x}_i,\mathrm{x}_j)\right\}, \tag{3.3}$$

where $\alpha_1, \alpha_2, \ldots, \alpha_n$ are nonzero Lagrange multipliers constrained to $0 \leq \alpha_i \leq \varsigma$, and $\Sigma_i\alpha_i y_i = 0$, for $i = 1, \ldots, n$. Some commonly implemented kernel functions are the linear kernel, the polynomial kernel, and the radial-basis-function (RBF) kernel [74].

$$K(\mathrm{x}_i,\mathrm{x}_j)=\exp\left(-\frac{\|\mathrm{x}_i-\mathrm{x}_j\|^2}{2\sigma^2}\right), \tag{3.4}$$

where σ is a width parameter of the RBF that controls the generalization capacity of the machine. The decision function is then given as

$$f(\mathrm{x})=\mathrm{sgn}\left(\sum_{i=1}^{n}y_i\alpha_i K(\mathrm{x}_i,\mathrm{x})+b\right). \tag{3.5}$$

SVM has become one of the most widely used nonparametric classifiers for hyperspectral data, in part due to its relative independence from data dimensionality, and it is included now in many software libraries.

3.4.2 Bayesian Parametric and Nonparametric Classification

Bayesian classification entails modeling class-conditional likelihoods [$p(\mathbf{x}|y_i)$, $i \in \{1, \ldots, c\}$] through an underlying probability model that is learned in the feature space (or some other appropriate subspace) from the training data. The state of the art in such methods for hyperspectral classification assumes that the class-conditional likelihoods are best modeled as mixtures (weighted linear combination) of "basis" Gaussian density functions. Depending upon whether one assumes there are finite or infinite Gaussian components in the mixture, the resulting model is either a traditional Gaussian mixture model or infinite Gaussian mixture model (IGMM), respectively.

3.4.2.1 Finite Gaussian Mixture Model

A GMM is a weighted linear combination of a finite number (K) of Gaussian components, such that the resulting likelihood function of $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{n}$ in $\mathbb{R}^d$ is

$$p(\mathbf{x})=\sum_{k=1}^{K}\alpha_k\mathcal{N}(\mathbf{x},\mu_k,\Sigma_k) \tag{3.6}$$

where

$$\mathcal{N}\left(\mathbf{x},\mu_k,\sum_k\right)=\frac{1}{(2\pi)^{d/2}\left|\sum_k\right|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\mu_k)^{\mathrm{T}}\sum_k^{-1}(\mathbf{x}-\mu_k)\right]. \tag{3.7}$$

GMMs are hence a parametric representation of the underlying probability model of the data, where the parametrization is formed by the parametric representations of each of the K Gaussian distributions (mean vector, μ_k, and covariance matrix, Σ_k), as well as the weights of each component in the mixture α_k, $\Theta=\{\alpha_k,\mu_k,\Sigma_k\}_{k=1}^{K}$.

Training GMM models (i.e., given $\mathbf{X}$, estimating $\Theta=\{\alpha_k,\mu_k,\Sigma_k\}_{k=1}^{K}$) is typically undertaken using the expectation maximization (EM) approach. A crucial design choice when training such GMM models is then determining the number of mixtures/components, K. This is often learned empirically by making use of an information theoretic measure, such as the Akaike information criterion (AIC) [75].

Such a Bayesian model can be used for both unsupervised and supervised learning. In an unsupervised learning framework, GMMs are often used as a robust Bayesian paradigm for clustering data—for instance, given an imagery data set (a set of pixels, $\mathbf{X}$) with no labels, cluster the data into its -K constituent clusters, where K is the underlying number of components (e.g., number of classes in the scene, assumed to be known a priori). In a supervised learning framework, GMMs can be used to learn class-conditional likelihood functions, with the assumption that each class may have a likelihood function that requires a mixture of multiple Gaussians to represent it effectively. In the context of remotely sensed imagery, such multimodal class-conditional distributions may arise due to practical effects (e.g., the same class in the image under different illumination conditions or variations in spectral response of vegetation due to varying vegetation stress within the same class in different parts of the scene).

3.4.2.2 Infinite Gaussian Mixture Models

A fundamental challenge of traditional GMMs is that the number of mixture components, K, may be difficult to ascertain a priori, and empirical approaches to estimating K may result in under- or overestimation of the required number of components within the mixture model. A recent development in the field of Bayesian nonparametrics, the IGMM, addresses this issue very effectively. Unlike traditional GMMs, the number of mixture components in IGMMs can be ascertained as part of the Bayesian inference process. IGMMs are a specific variant [76] of Dirichlet process mixture models (DPMMs). In DPMM formulations, often a stick-breaking construction process [76] is used to generate mixture/component weights that have a Dirichlet process prior placed on them. Unlike traditional GMMs, IGMMs do not assume that the number of mixture components is known a priori and instead infer the mixture components via Bayesian inference strategies, such as Markov chain Monte Carlo sampling [76].

3.4.2.3 Practical Issues: Dimensionality

A key point to be made here with respect to Bayesian inference approaches is that the success of such approaches is contingent on the training sample size relative to the dimensionality of the feature space. For example, with a K-component GMM, assuming full covariance matrices, one needs to estimate $K(1+d(d-1)/2)+Kd$ parameters from the training data. Learning such high-dimensional parameters with limited training data is highly impractical—hence, feature extraction (dimensionality reduction) is often a critical preprocessing step to such Bayesian inference strategies. Figure 3.9 illustrates the performance of GMM-based classification coupled with a locality preserving feature reduction approach for classification of the Indian Pine 1992 data set. For this data set, LFDA-SVM and LFDA-GMM consistently resulted in the highest overall accuracy. In our recent work [77], we found that for hyperspectral image analysis, including for vegetation, locality-preserving approaches

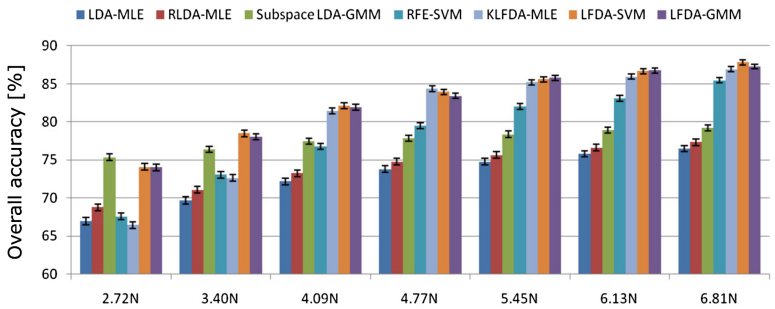


FIGURE 3.9 Classification accuracy as a function of training sample size using GMM classifiers coupled with LFDA-based feature reduction for the Indian Pine 1992 data set. Comparisons to baseline methods are also provided. Linear discriminant analysis-max likelihood estimate (LDA-MLE); regularized linear discriminant analysis-max likelihood estimate (RLDA-MLE); subspace LDA-Gaussian mixture model (Subspace LDA-GMM); recursive feature elimination-SVM (RFE-SVM); kernel local Fisher discriminant analysis-SVM (KLFDA-SVM); local Fisher discriminant analysis-SVM (LFDA-SVM); local Fisher discriminant analysis-Gaussian mixture models. (Adapted from W. Li et al. *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 4, pp. 1185–1198, 2012. [78])

to subspace learning are the most effective feature extraction/preprocessing for Bayesian inference, as they preserve locality (neighborhood structure) of the feature space in the embedded subspace.

3.4.3 Deep Learning of Hyperspectral Imagery Data

Deep learning has emerged as a very effective approach in contemporary computer vision and signal analysis tasks. This has been driven in part by increased availability of high-performance computing infrastructures and rich libraries for training models. Deep learning algorithms include deep convolutional neural networks (CNNs), recurrent neural networks (RNNs), and deep belief networks (DBNs) that have been successfully deployed for speech recognition, computer vision, natural language processing, and, more recently, remote sensing applications [79–81]. CNNs and their variants have been successfully used for tasks such as large-scale object detection, transfer learning/domain adaptation, and so on. RNNs and their variants have also been demonstrated to be useful for temporal modeling for applications such as speech recognition [82,83].

Within the context of pixel-level hyperspectral image analysis, we have observed that the spectral reflectance characteristics of the data can be modeled quite effectively through what we call convolutional recurrent neural networks (CRNNs). The approach entails a series of convolutional and pooling layers followed by recurrent neural network layers, as shown in Figure 3.10. The convolutional layers are adept at extracting stable, locally invariant features from the spectral reflectance features. Pooling layers help minimize effects of overfitting. The recurrent layers extract the interchannel relationship in spectral reflectance profiles. Each convolutional layer has multiple 1-D convolutional filters where the filter support is a data-dependent parameter. The pooling layers are used for subsampling to reduce the dimensionality of the network, which can help reduce computation and control overfitting. A recurrent layer has a recursive function f that takes as input one input vector $\boldsymbol{x}_t$ and the previous hidden state $\boldsymbol{h}_{t-1}$, and returns the new hidden state as:

$$h_t = f(x_t, \boldsymbol{h}_{t-1}) = tanh(\boldsymbol{W}\boldsymbol{x}_t + \boldsymbol{U}\boldsymbol{h}_{t-1}) \tag{3.8}$$

and the outputs are calculated as:

$$\boldsymbol{o}_t = softmax(\boldsymbol{V}\boldsymbol{h}_t) \tag{3.9}$$

where $\boldsymbol{W}$, $\boldsymbol{U}$, and $\boldsymbol{V}$ are the weight matrices that are shared across all steps, and the activation function *tanh* is the hyperbolic tangent function. In our framework (CRNN), the last hidden state (node)

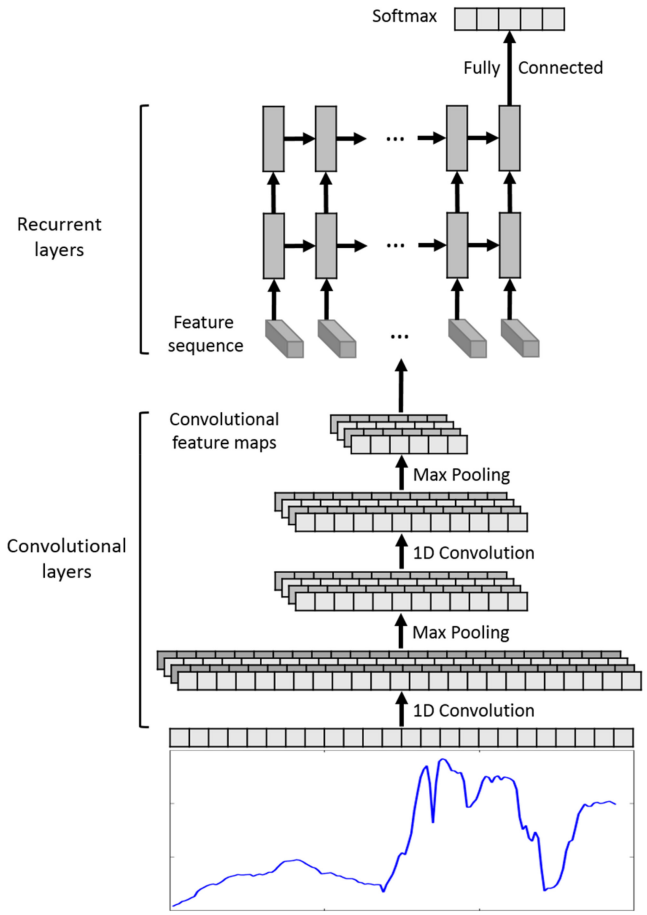


FIGURE 3.10 CRNN architecture for hyperspectral data classification. (Adapted from H. Wu and S. Prasad. *Remote Sensing*, vol. 9, no. 3, p. 298, 2017. [86])

of the RNN is fully connected to the classification layer that is based on the softmax activation function. Cross-entropy is used as a loss function, and mini-batch gradient descent is used to find the best network parameters. Gradients in the convolutional layers are calculated by traditional back-propagation, while the gradients in the recurrent layers are computed by back-propagation through time (BPTT) [84]. As an illustration of the potential for pixel-level deep learning based on CRNN for vegetation classification, classification maps with the Indian Pine 2010 data set as well as comparisons to baselines are provided in Table 3.3.

3.4.4 Segmentation-Based Approaches to Support Image Classification

Segmentation is often an effective precursor to classification, providing the capability to identify homogeneous regions that are classified as objects. A myriad of image segmentation approaches have been proposed and developed over the years (see, e.g., [87]). These approaches can be adapted to segmenting hyperspectral imagery as a preprocessing stage to classification by first reducing the data dimensionality through feature selection and extraction (see previous section) or by utilizing an appropriate region similarity (or dissimilarity) criterion.

A dissimilarity criterion designed for hyperspectral data is the spectral angle mapper (SAM) criterion [88]. An important property of the SAM criterion is that poorly illuminated and more

TABLE 3.3

Classification Results (Overall Accuracy and Standard Deviations) Obtained by Different Approaches with Different Numbers of Training Samples on the Indian Pine 2010 Data Set

# Training Samples	1900	3800	5700
RBF-SVM	92.82(±1.07)	94.36(±0.93)	95.13(±0.64)
CNN	93.11(±0.95)	94.53(±0.39)	95.84(±0.31)
RNN	84.83(±1.62)	89.74(±0.98)	91.86(±0.77)
CRNN	94.43(±1.01)	96.24(±0.60)	96.83(±0.47)

Source: Adapted from H. Wu and S. Prasad. *Remote Sensing*, vol. 9, no. 3, p. 298, 2017. [86]

brightly illuminated pixels of the same color will be mapped to the same spectral angle despite the difference in illumination.

The most successful approaches to spatial-spectral image segmentation are based on best merge region growing. An early version of best merge region growing, hierarchical step-wise optimization (HSWO), is an iterative form of region growing in which the iterations consist of finding the most optimal or best segmentation with one region less than the current segmentation [89]. The HSWO approach can be summarized as follows:

1. Initialize the segmentation by assigning each image pixel a region label. If a presegmentation is provided, label each image pixel according to the presegmentation. Otherwise, label each image pixel as a separate region.
2. Calculate the dissimilarity criterion value, d, between all pairs of spatially adjacent regions, find the smallest dissimilarity criterion value, T_{merge}, and merge all pairs of regions with $d = T_{merge}$.
3. Stop if no more merges are required. Otherwise, return to step 2.

HSWO naturally produces a segmentation hierarchy consisting of the entire sequence of segmentations. For practical applications, however, a subset of segmentation needs to be selected from this exhaustive segmentation hierarchy. A portion of such a segmentation hierarchy is illustrated in Figure 3.11 (the selection of a single segmentation from a segmentation hierarchy is discussed in a later section).

A unique feature of the segmentation hierarchy produced by HSWO and related region growing segmentation approaches is that the segment or region boundaries are maintained at the full image spatial resolution for all levels of the segmentation hierarchy. This is important for classification problems.

Many variations on best merge region growing have been described in the literature. As early as 1994, [90] described an implementation of the HSWO form of best merge region growing that utilized a heap data structure [91] for efficient determination of best merges and a dissimilarity criterion based on minimizing the mean squared error between the region mean image and original image. The main differences between most of these region growing approaches are the dissimilarity criterion employed and, perhaps, some control logic designed to remove small regions or otherwise tailor the segmentation output. In complex scenes, such as remotely sensed images of the Earth, objects with similar spectral signatures (e.g., lakes, agricultural fields, buildings, etc.) appear in spatially separated locations. In such cases, it is useful to aggregate these spectrally similar but spatially disjoint region objects into groups of region objects, or region classes. This is the basis of the hybridization of HSWO best merge region growing with spectral clustering [92,93] called HSeg (hierarchical segmentation). HSeg adds to HSWO a step following each step of adjacent region merges in which all pairs of spatially nonadjacent regions are merged that have $dissimilarity \leq S_w T_{merge}$,

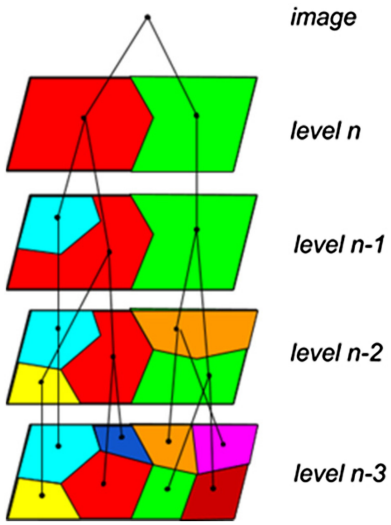


FIGURE 3.11 The last four levels of an *n*-level segmentation hierarchy produced by a region growing segmentation process. Note that when depicted in this manner, the region growing process is a "bottom-up" approach. (Adapted from J. C. Tilton et al. "Image segmentation algorithms for land categorization," in *Remotely Sensed Data Characterization, Classification, and Accuracies*, 2015, pp. 317–342. [87])

where $0.0 \leq S_w \leq 1.0$ is a factor that sets the priority between spatially adjacent and nonadjacent region merges. Note that when $S_w = 0.0$, HSeg reduces to HSWO.

A recursive divide-and-conquer approximation of HSeg (called RHSeg) [94] was developed to enable a straightforward parallel implementation. The computational requirements of HSeg were further reduced by refinements discussed in [93]. HSeg segmentation results are also compared in [93] with other segmentation approaches based on region growing.

Determining the optimal level of segmentation detail for a particular application is a challenge for all image segmentation approaches. The level of segmentation detail produced by a segmentation approach can usually be specified by adjusting one or more of the approach's parameters. Additionally, segmentation approaches based on region growing can also be easily designed to output a hierarchical set of image segmentations that can be examined later to select an optimal level of segmentation detail. In the case of HSeg, a particular set of hierarchical segmentations can be specified by providing a set of merge thresholds or number of region classes. HSeg can also automatically select a hierarchical set of image segmentations over a specified range of the number of region classes. This automatic approach outputs the image segmentation result at the region growing iteration prior to the point where any region classes are involved in a second merge since the last segmentation results output.

The following subsections describe various approaches for selecting an optimal level of segmentation detail out of an image segmentation hierarchy, such as that produced by HSeg:

1. *A Semisupervised Approach for Selecting Segmentations from a Segmentation Hierarchy*: [95] describes a semisupervised approach for adaptively selecting segmentations from the region class segmentation hierarchy produced by HSeg (or RHSeg). With the HSegLearn tool described therein, an analyst selects region classes (groups of region objects) as positive or negative examples of a specific ground cover class. Based on these selections, HSegLearn searches the HSeg segmentation hierarchy for the coarsest level of segmentation at which the submitted positive example region classes do not conflict with the submitted negative example region classes and displays the submitted positive example region classes at

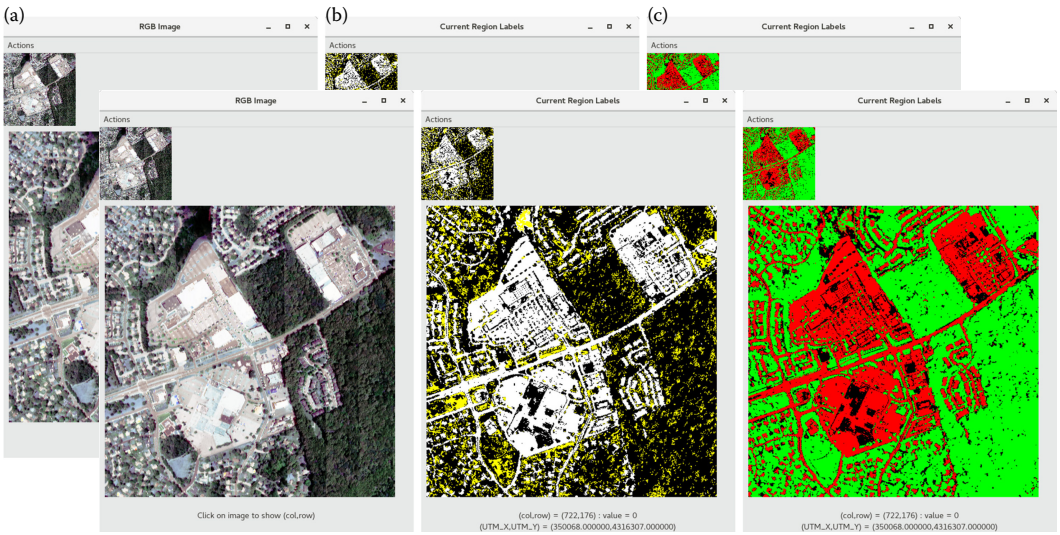


FIGURE 3.12 Example of using HSegLearn to generate a ground reference map of vegetation. (a) RGB image panel displaying a subsection of a hyperspectral image over a portion of downtown Bowie, MD. (b) HSegLearn Current Region Labels Panel after selecting some positive example region classes (vegetation, yellow) and some negative example region classes (nonvegetation, white). (c) HSegLearn Current Region Labels Panel: the positive example regions (vegetation, green) and negative regions (nonvegetation, red); region labeling by finding the coarsest levels in the HSeg segmentation hierarchy that do not conflict with the analyst labeling. These hyperspectral data were obtained by NASA Goddard's LiDAR, Hyperspectral and Thermal (G-LiHT) Airborne Imager on June 28, 2012, at a nominal 1-m spatial resolution. The data set has 114 spectral bands over the spectral range of 418–918 nm.

that level of segmentation detail. HSegLearn was successfully used to generate ground reference data from 1- to 2-m resolution satellite imagery (mainly WorldView 2 imagery) for training of the classification of Landsat TM and ETM+ data [96] for a global mapping of human-made impervious surfaces. This tool can also be used with hyperspectral data, as demonstrated by the example shown in Figure 3.12.

2. *Selecting Segmentations from a Segmentation Hierarchy by Analyzing Classification Error Rates*: [97] developed an approach for selecting the best level of segmentation detail from the segmentation hierarchy produced by HSeg (or RHSeg) for fusion of the region object labeling from HSeg with a pixel-based random forest classification of Landsat TM or ETM+ data covering the North American continent. This approach was used to determine the merge threshold for the best level of segmentation detail in the HSeg segmentation in each agricultural zone. This was done by analyzing the error rates across all the segmentation hierarchies in all image tiles containing reference objects. The merge threshold was determined as the threshold for the segmentation hierarchy with the lowest error rate in a stable portion of a graph of error rate versus segmentation hierarchy level.
3. *Selecting Segmentations from a Segmentation Hierarchy Utilizing Active Learning*: [98] describes an approach utilizing an active learning framework for selecting the best segmentation obtained by pruning the segmentation hierarchy produced by HSeg from a hyperspectral image. By pruning we mean adaptively selecting the level of segmentation detail from varying levels of the segmentation hierarchy throughout the image. Since this approach utilizes active learning, we defer discussion of this approach to the section on active learning.

3.5 CHALLENGES AND ADVANCED APPROACHES FOR CLASSIFICATION OF VEGETATION

3.5.1 Multisource/Multitemporal/Multiscale Challenges and Approaches

3.5.1.1 Multiple Kernel Learning

It is well understood that with traditional SVM classifiers, the choice of kernel function and the associated kernel parameter (e.g., the width of the RBF kernel) have a significant impact on the classification performance. A key practical aspect of using SVM classifiers with hyperspectral data is the need to "tune" the kernel parameters to find the appropriate (data-dependent) parameters that result in optimal decision boundaries. Additionally, when dealing with multisource data (e.g., data from different sensors), traditional single-kernel SVMs do not offer a compelling approach to fuse such data into a unified classification product. In recent work, multikernel learning methods have been shown to fill in these gaps that traditional SVMs struggle to classify. Multikernel learning can be thought of as learning a traditional SVM, with the exception that instead of traditional Mercer's kernels, one uses a mixture of "basis" kernels, where the mixing weights are learned as part of the SVM optimization.

In the multisource scenario, for a specific source p, the combined kernel function K between two pixels $\mathbf{x}_i^p$ and $\mathbf{x}_j^p$ can be represented as

$$K(\mathbf{x}_i^p, \mathbf{x}_j^p) = \sum_{m=1}^{M} d_m K_m(\mathbf{x}_i^p, \mathbf{x}_j^p) \quad \text{s.t. } d_m \geq 0, \text{ and } \sum_{m=1}^{M} d_m = 1, \tag{3.10}$$

where M is the number of candidate basis kernels representing different kernel parameters, K_m is the mth basis kernel, and d_m is the associated weight for it. The SimpleMKL [99] algorithm is an implementation of the MKL framework, where the optimization problem is posed as:

$$\min_d J(d), \quad \text{s.t.} \, d_m \geq 0, \text{and} \sum_{m=1}^{M} d_m = 1$$

$$J(d) = \begin{cases} \min_{\mathbf{w},b,\xi} \dfrac{1}{2} \displaystyle\sum_{m=1}^{M} \frac{1}{d_m} \|\mathbf{w}_m\|^2 + C \sum_{i=1}^{N} \xi_i \\ \text{s.t. } y_i \left(\displaystyle\sum_{m=1}^{M} \left\langle \mathbf{w}_m, \Phi_m(\mathbf{x}_i^p) \right\rangle + b \right) \geq 1 - \xi_i \\ \xi_i \geq 0, \quad \forall i = 1, 2, \ldots, N, \end{cases} \tag{3.11}$$

where $\Phi_m(\mathbf{x}^p)$ is the kernel mapping function of $\mathbf{x}_i^p$, $\mathbf{w}_m$ is the weight vector of the mth decision hyperplane, C is the regularization parameter controlling the generalization capabilities of the classifier, and ξ_i is a positive slack variable.

As with traditional SVMs, SimpleMKL also has a dual representation:

$$\max\left\{L(\alpha_i,\alpha_j)=\sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \sum_{m=1}^{M} d_m K_m(\mathbf{x}_i^p,\mathbf{x}_j^p)\right\}$$
$$\text{s.t.}\begin{cases}\displaystyle\sum_{i=1}^{N}\alpha_i y_i = 0\\ \alpha_i,\alpha_j\in[0,C], \quad \forall i,j=1,2,\ldots,N\\ d_m\geq 0, \text{ and } \displaystyle\sum_{m=1}^{M} d_m = 1,\end{cases} \tag{3.12}$$

where α_i and α_j are Lagrange multipliers. Kernel weights d_m are learned via gradient descent by updating it along the direction of the negative gradient of $L(\alpha_i, \alpha_j)$. The outcome of SimpleMKL yields a predicted label per test sample, and by using Platt's approach, a class-conditional posterior probability $P(y = 1|x)$ is estimated, obtaining the posterior probability that allows quantification of uncertainty associated with the underlying classification.

An alternative approach to use the MKL framework entails learning source-specific optimal MKL classifiers, which are then fused through decision fusion strategies, such as majority voting or linear opinion pools. In this alternative construction, the mixture of kernels per SVM yields a much more discriminative and linearly separable Hilbert space than a traditional single kernel–based SVM. We used multikernel SVMs to systematically fuse multisensor (hyperspectral and LiDAR) geospatial data in an active learning paradigm [100].

3.5.1.2 Transfer Learning and Domain Adaptation

A common image analysis scenario when dealing with multisource data involves transferring knowledge/information between sources. For instance, given a source data set rich in labeled data where one can optimally learn a robust classifier, can this knowledge/model be transferred to a (different) target data set where there may not otherwise be sufficient training data to train classifiers for effective classification of target data? It is common to encounter such scenarios when undertaking multiscale, multisensor, and multitemporal data sets. Here, we refer to recent efforts [53,101,102] aimed at transferring knowledge across sources (e.g., across different sensors or different viewpoints or times corresponding to the same sensor).

Denote the source domain as $\mathcal{D}_S$ with samples $\{x_1, \ldots, x_n\} \in \mathbb{R}^{d_s}$ and corresponding class labels $\{l_{x_1},\ldots,l_{x_n}\}$. Similarly, denote the target domain as $\mathcal{D}_T$ with labeled samples $\{y_1,\ldots,y_m\} \in \mathbb{R}^{d_t}$, $d_s \neq d$t, and corresponding class labels $\{l_{y_1},\ldots,l_{y_m}\}, m \ll n$. Given vectors $x\in\mathcal{D}_S$ and $y\in\mathcal{D}_T$, the goal is to project these data to a latent space $\mathbb{R}^{d_c}$ that is discriminative for the underlying classification task.

In recent work, we developed mappings **A** and **AB** that maximize the overlap of within-class samples in the latent space by ensuring that within-class samples in $\mathcal{D}_S$ and $\mathcal{D}_T$ are located in the same cluster/region of the latent feature space. Table 3.4 depicts results of this domain adaptation based on transformation learning (DATL) approach with the Botswana data and compares it to traditional approaches (including semisupervised transfer component analysis (SSTCA) and kernel principal component analysis) to domain adaptation. Class specific accuracies reported in Table 3.5 result from training on a large pool of source data (May Botswana Hyperion image) and transferring

TABLE 3.4

Overall KNN Classification Accuracies (%) and Standard Deviations (in Parentheses) for Botswana Data Set from May (Source) and July (Target)

	# Target Domain Training Samples per Class							
Algorithm	**1**	**3**	**5**	**7**	**9**	**11**	**13**	**15**
KPCA	34.0 (4.5)	35.1 (4.6)	37.5 (4.8)	38.5 (4.5)	40.5 (4.6)	40.8 (4.8)	41.2 (4.6)	42.0 (5.5)
SSTCA	47.7 (8.4)	47.5 (7.2)	50.5 (9.1)	49.6 (7.6)	52.4 (7.5)	53.0 (7.8)	52.6 (7.3)	51.7 (6.5)
DATL	46.6 (12.2)	63.6 (13.4)	68.2 (8.5)	71.3 (10.0)	73.3 (7.9)	76.9 (5.9)	76.8 (7.9)	77.6 (6.7)

Source: Adapted from X. Zhou and S. Prasad. *IEEE Transactions on Computational Imaging*, vol. 3, no. 4, pp. 822–836, December 2017. [102]

TABLE 3.5

Class-Dependent Accuracies (%) for Botswana Data from May (Source) and July (Target) with Five Training Samples per Class from the Target Domain

	Class Index								
Algorithm	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
KPCA	100.0	13.4	50.1	97.9	0.6	14.0	0.5	1.8	15.1
SSTCA	100.0	54.4	61.1	99.8	10.0	32.0	15.4	18.1	37.8
DATL	68.5	50.4	47.2	94.4	77.3	42.8	59.2	89.5	84.5

TABLE 3.6

Overall Classification Accuracies (%) and Standard Deviations (in Parentheses) for the Aerial and Street View Wetland Hyperspectral Data Set

	# Target Domain Training Samples per Class							
Algorithm	**1**	**3**	**5**	**7**	**9**	**11**	**13**	**15**
KPCA	26.4 (7.6)	46.8 (9.5)	59.9 (3.3)	64.4 (3.4)	67.0 (5.0)	70.0 (5.3)	72.5 (5.6)	73.6 (6.1)
SSTCA	25.4 (9.8)	51.0 (8.3)	65.1 (3.2)	70.0 (3.0)	74.3 (2.4)	77.9 (3.2)	80.9 (3.5)	82.7 (4.2)
DATL	42.0 (5.4)	59.8 (6.6)	68.2 (7.8)	70.6 (8.0)	74.5 (7.0)	75.6 (4.7)	78.7 (5.9)	81.0 (4.9)

Source: Adapted from X. Zhou and S. Prasad. *IEEE Transactions on Computational Imaging*, vol. 3, no. 4, pp. 822–836, December 2017. [102]

the models learned to classify a different image acquired over the same region in July (target image) using between 1 and 15 target training samples per class to facilitate the alignment.

Accuracies in Tables 3.6 and 3.7 represent the coastal wetland ecosystem monitoring application where the source domain refers to street-view (side-looking, terrestrial) hyperspectral imagery captured during our field campaigns to identify the vegetation cover. The target domain refers to aerial imagery.

The goal of such a framework is to learn models from the rich source domain data (due to being imaged at close range, there is an abundance of pixels per class to train models in the source domain) and transfer this knowledge to undertake classification in the target domain. For this data, as few as 5–15 samples per class from the target domain, very effective classification accuracies are achieved leveraging domain information. Results are particularly promising because of different

TABLE 3.7

Class-Dependent KNN Classification Accuracies (%) for the Aerial and Street View Wetland Hyperspectral Data Set with Five Training Samples per Class from the Target Domain (Aerial View)

	Class Index											
Algorithm	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
KPCA	64.7	48.5	57.5	29.7	67.5	23.5	45.8	72.5	88.0	78.7	75.3	67.3
SSTCA	65.0	63.3	59.8	30.8	75.0	32.7	49.2	73.2	91.8	86.5	81.2	72.5
DATL	71.0	66.5	69.3	41.6	67.4	25.2	47.5	71.5	82.0	91.2	70.6	62.6

Source: Adapted from X. Zhou and S. Prasad. *IEEE Transactions on Computational Imaging*, vol. 3, no. 4, pp. 822–836, December 2017. [102]

sun-sensor-canopy geometry but also different sensors (source spans the VNIR range, while the target spans the VNIR-SWIR range).

3.5.2 Limited Training Samples: Exploiting Unlabeled Samples

3.5.2.1 Semi-Supervised and Active Learning

Semisupervised learning (SSL) approaches, where unlabeled samples are classified and some are incorporated into the training set, have been investigated by many researchers for classification of high-dimensional data. SSL directly utilizes the unlabeled data to facilitate the learning process without requiring any human-labeling efforts. In addition to semisupervised learning approaches, many researchers have explored active learning (AL) as an alternative strategy for leveraging unlabeled samples to mitigate the impact of limited training samples for supervised classification of remotely sensed hyperspectral data. Unlike SSL approaches, AL heuristics focus on identifying the most "informative" unlabeled samples, then obtaining the corresponding labels and incorporating the newly labeled data into the training pool.

In traditional supervised classification, as shown in the box in Figure 3.13, labeled samples/features are presented to the classifier for training, and unlabeled samples from the data pool are then classified using the resulting model. AL involves an iterative training/evaluation phase, whereby an initial model is developed using a small number of samples, and the classification results are evaluated. Samples are then selected from the unlabeled pool, evaluated according to "query" criteria, and ranked. Sample(s) are chosen from the list for labeling according to specified selection criteria, labeled, and incorporated

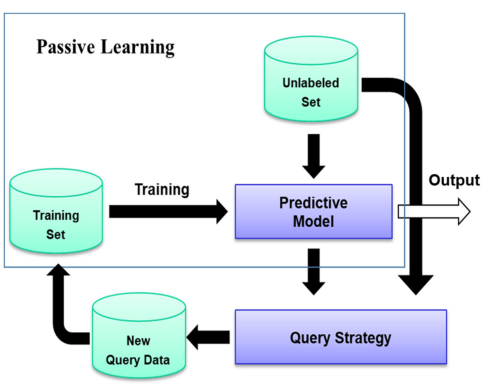


FIGURE 3.13 Flow chart for active and passive learning.

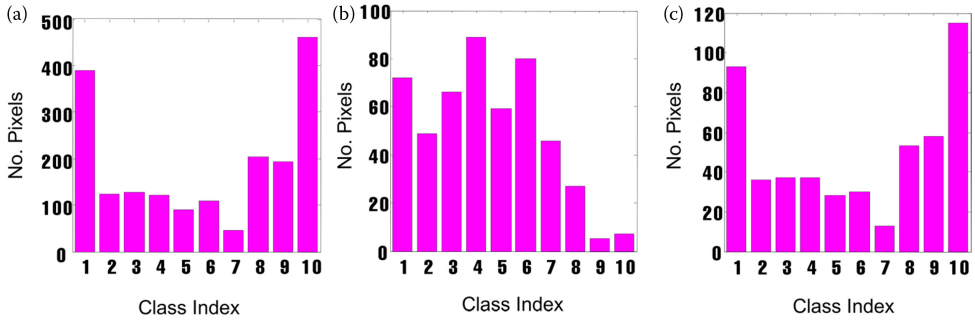


FIGURE 3.14 Distribution of labeled class samples for Kennedy Space Center data. (a) True distribution, (b) AL sampling, (c) random sampling.

into the training set. The remaining candidates are returned to the unlabeled pool. The classifier is updated using the augmented training set, and the process is repeated until a stopping criterion (e.g., number of iterations, number of samples, classification accuracy) is achieved.

AL strategies typically result in biased sampling of classes, with more samples being selected from classes that are more difficult to discriminate, as shown in the example of Figure 3.14. AL strategies are generally characterized by the criteria used to select candidate samples for labeling. The approaches include (1) margin sampling (MS)–based approaches, where the samples closest to the separating hyperplane of the classifier, such as support vector machines, are considered as the most uncertain ones [103,104]; (2) committee of learners–based approaches, where those samples exhibiting maximal disagreement between the different classification models are selected [8,105,106]; and (3) class probability distribution–based approaches, where breaking ties (BT) is a representative method in which the difference between the two highest posterior probabilities is used to quantify the uncertainty of a pixel [107,108]. Algorithms are typically implemented with spectral data as the single-source inputs, although multiple-source inputs have also been studied [109,110]. Active learning for classification has been used in multiple applications [111,112] including large-scale scenarios [113]. We refer to [9,114] for surveys on AL methods.

Results from [106] are included to illustrate AL based on a committee of learners. Views can be composed of multiple classifiers resulting from different methods, different inputs, and so on. Here, multiple views of the problem were derived by segmenting the spectral data into disjoint contiguous subband sets generated by (1) correlation of contiguous bands, (2) *k*-means–based band clustering, (3) deterministic selection of every *k*th band (band slicing), and (4) random sampling. Views generated by correlation and clustering are diverse, but may differ in their discriminative ability for individual classes, so there is a risk of insufficiency for the classification, while views obtained from band slicing may be redundant, but are sufficient for covering the full space of inputs. Finally, random sampling provides diverse views, although they are not guaranteed to be either sufficient or accurate. Figure 3.15 illustrates multiview subsetting of the input space of the KSC data based on interband correlation. A two-stage query and sample selection process was used, where the first ranked samples by the maximum disagreement of the classification results across views, and the second invoked an entropy criterion to increase diversity of the samples. Figures 3.16 and 3.17 show the accuracy curves as AL progresses. Margin sampling had the highest overall accuracy, although multiview methods converged to approximately the same overall accuracy.

3.5.2.2 Segmentation-Based Active Learning

Most of the AL methods proposed in the literature deal solely with spectral information, but improvements in terms of classification accuracies can be obtained by also exploiting the spatial dimension. Only recently, researchers have started to integrate spatial information with spectral features in the AL framework [109,115–117]. A recent strategy [98] that incorporates AL,

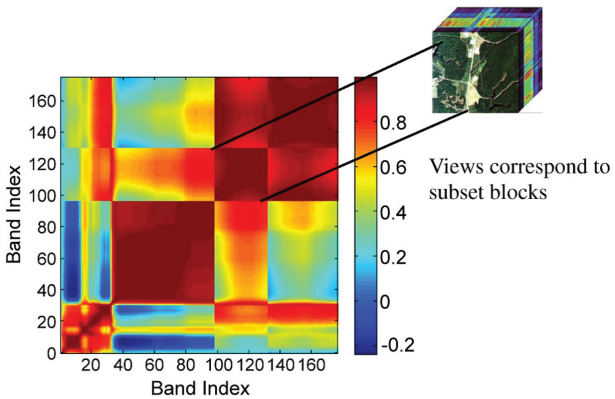


FIGURE 3.15 Correlation matrix for Kennedy Space Center AVIRIS data. Each block corresponds to a view.

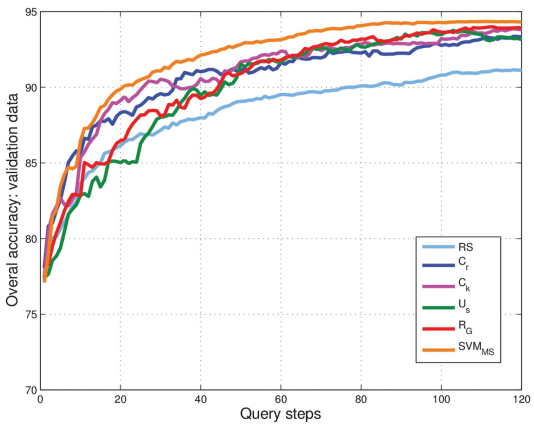


FIGURE 3.16 Overall classification accuracy for SVM classification of Kennedy Space Center with multiview active learning based on band correlation C_r, k-means clustering C_k, uniform band slicing U_s, and random view generation R_U compared to random sampling RS and margin sampling SVM_{MS}. (Adapted from W. Di and M. M. Crawford, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 5, pp. 1942–1954, 2012. [106])

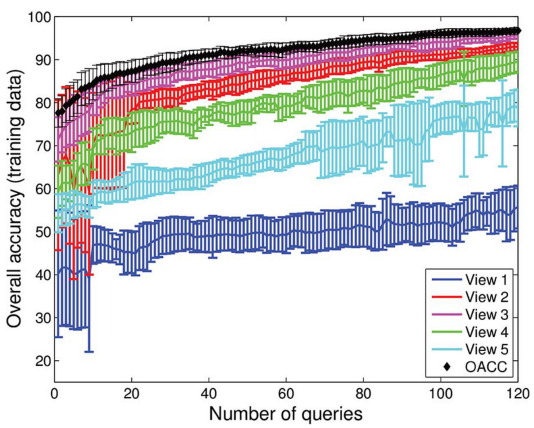


FIGURE 3.17 The overall classification accuracy and view performance derived from correlation-based view generation C_r for Kennedy Space Center data. (Adapted from W. Di and M. M. Crawford, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 5, pp. 1942–1954, 2012. [106])

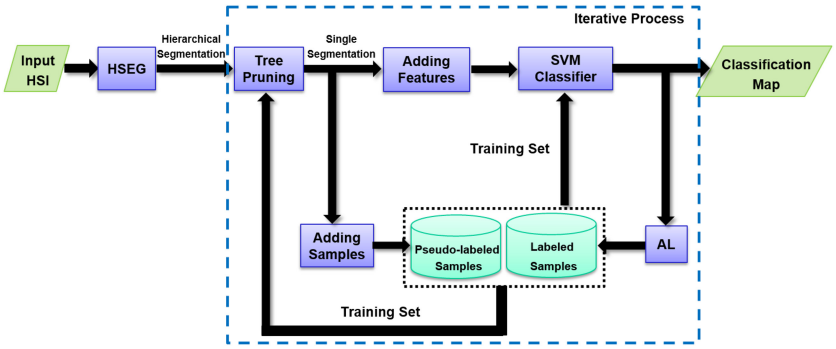


FIGURE 3.18 Flow chart illustrating the segmentation-based AL framework that incorporates AL, semisupervised learning, and segmentation into a unique framework. (Adapted from Z. Zhang et al. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 2, pp. 640–654, 2016. [98])

semisupervised learning, and segmentation into a unique framework is shown in Figure 3.18 and detailed in the following.

The approach relies on the HSeg algorithm introduced earlier in this chapter, whose output is a segmentation hierarchy that contains different levels of detail of the image. The appropriate level can be selected in a supervised way by quantitatively evaluating the segmentation results at each hierarchical level (as discussed previously in this chapter and in [97]) or in an unsupervised way by interactively using the HSegViewer tool. Both approaches assume that the best segmentation corresponds to a single specific level of the hierarchy. However, coherent objects may be found at different levels [118] and therefore, the best segmentation should be defined by selecting regions at different levels. This process is usually referred to as pruning, in which subtrees of the hierarchy that are homogeneous with respect to some defined homogeneity criteria are removed. Although different pruning strategies based on supervised [119], semisupervised [120], and unsupervised [121] strategies have been proposed, they are not suitable to be used within an AL framework. While AL and SSL have different workflows, they both aim to make the best use of unlabeled data and reduce human labeling efforts. It is natural to combine the two strategies to take advantage of both paradigms in the classification task. The main idea of the strategy is to find an optimal segmentation map from the segmentation hierarchy by considering a novel supervised pruning strategy. This strategy aims at removing redundant subtrees composed of nodes that are homogeneous with respect to some criteria of interest from the HSeg output. As a result, it generates an optimal segmentation map that can provide spatial information for the classification. The best segmentation does not represent one of the actual levels of the hierarchy, but incorporates regions selected from potentially different levels. Two merging criteria based on the node size and the Bhattacharyya coefficient are considered. The whole pruning process is integrated within the AL framework. Compared to the unsupervised pruning strategy proposed in [121], the new method can exploit the labeled information provided by the user (which increases through the AL process) and thus update the best segmentation map at each AL iteration. At the end of the pruning, a single segmentation map is obtained, which is further exploited to incorporate spatial information into the framework in two different ways: (1) spatial features (e.g., mean and standard deviation) are extracted from each segment and concatenated with the original spectral ones into a single stacked feature vector. An SVM is adopted as the back-end classifier and trained on the enriched feature space. (2) The continuously updated segmentation map is considered to expand the training set by employing both AL and self-learning–based SSL approaches. At each iteration, both labeled and pseudo-labeled samples are added to the training set and used jointly to train the classifier. The most uncertain samples are selected using the BT criterion and then labeled by the human expert. Pseudo-labels are assigned automatically by taking advantage

of spatial information. Such pseudo-labeled samples help increase the size of the training set when the available labeled training set is small. For this purpose, first, the set of candidate unlabeled samples is defined as the samples that belong to regions with identically labeled samples based on the current segmentation map. The framework is validated experimentally on the Indian Pine 1992 data set. Improvements in terms of overall classification accuracies are exhibited by the new framework in comparison with other spectral-spatial strategies, as reported in Figures 3.19 through 3.21.

3.5.2.3 Active Metric Learning

Feature extraction and AL problems for hyperspectral image classification have usually been investigated independently. Considering a traditional AL-based hyperspectral image classification chain, feature extraction is usually executed first in the original high-dimensional feature space as a preprocessing step to obtain an optimally reduced feature space. This can be accomplished in an unsupervised way using manifold learning strategies or in a supervised way by exploiting the limited labeled information available at the beginning of the process. An AL algorithm is then applied in the reduced feature space to increase the number of points in the training set. However, the feature space extracted earlier may be suboptimal relative to the resulting training set. The unsupervised feature extraction step lacks connection with the classification problem or is performed using the few potentially nonrepresentative initial labeled samples. In both cases, the extracted feature space is fixed and does not interact in any way with the additional information provided by the user during the AL process. Therefore, even an optimal AL strategy cannot guarantee maximization of the classification accuracy since it is applied to a suboptimal feature space.

Novel solutions have recently been proposed in the literature with the aim of combining feature extraction and AL into a unique framework [122–124], as summarized in the flowchart of Figure 3.22. The overall idea is to learn and update a reduced feature space in a supervised way at each iteration of the AL process, thus exploiting the increasing labeled information provided by the user. In particular, the computation of the reduced feature space is based on the large-margin nearest

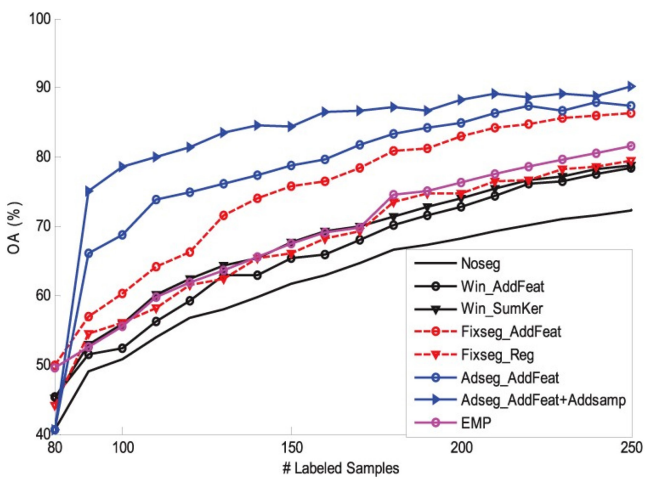


FIGURE 3.19 Learning curve (overall accuracy vs. number of labeled samples) on the Indian Pine 1992 data set. The methods *Adseg_AddFeat* and *Adseg_AddFeat+Addsamp* are compared with other spectral-spatial strategies. *Win_AddFeat*: spectral and spatial features extracted from a 3×3 window; *Win_SumKer*: spectral and spatial features extracted from a 3×3 window with a kernel-composite approach classifier; *Fixseg_AddFeat*: spatial features extracted from a fixed segmentation map; *Fixseg_Reg*: the fixed segmentation map is used as regularizer; *EMP*: morphological features extracted from the first two PCs. All methods adopt BT as the AL query criterion. (Adapted from Z. Zhang et al. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 2, pp. 640–654, 2016. [98])

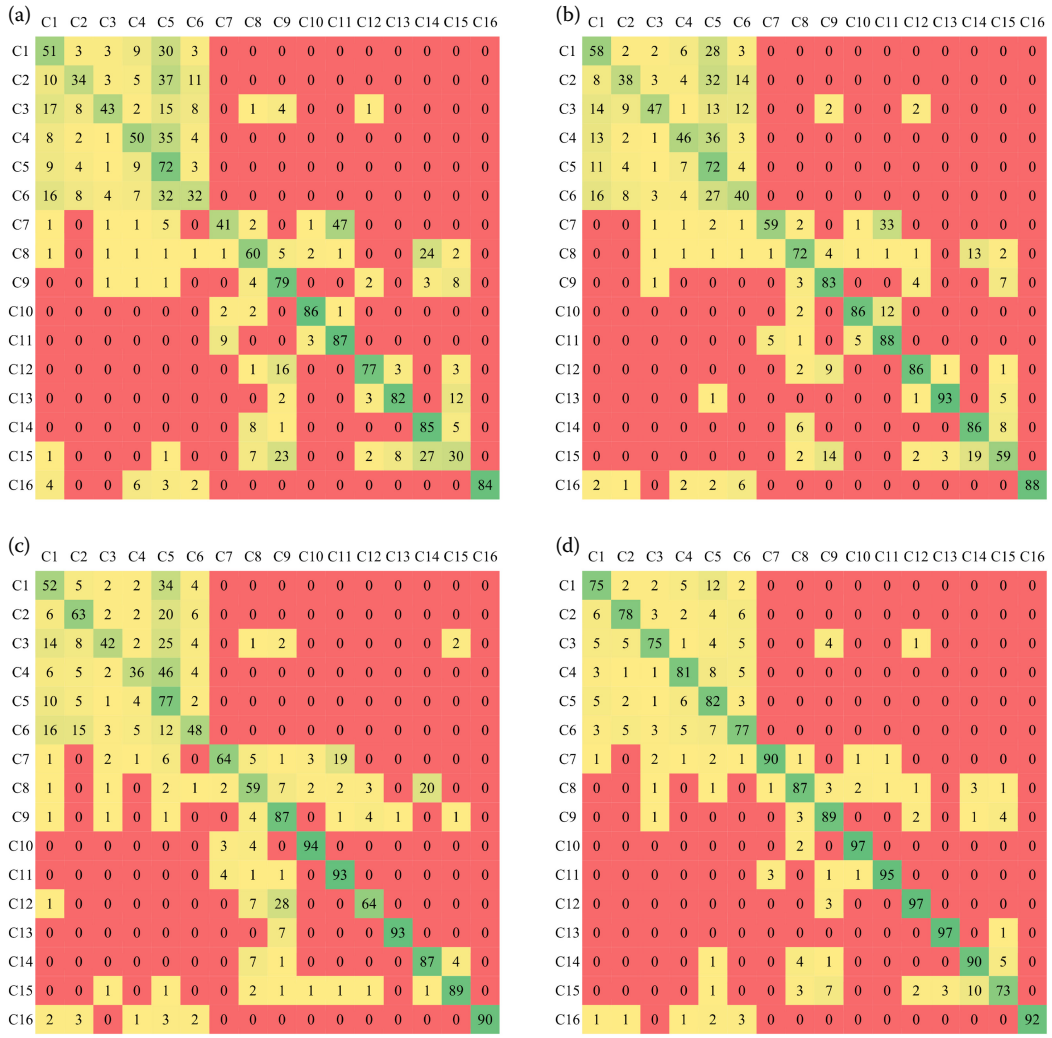


FIGURE 3.20 Normalized confusion matrices (values in percentage) achieved on the Indian Pine 1992 data set. (a) *Noseg*, (b) *Win_AddFeat*, (c) *EMP*, (d) *Adseg_AddFeat* + *AddSamp*. (Adapted from Z. Zhang et al. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 2, pp. 640–654, 2016. [98])

neighbor (LMNN) metric learning principle [125]. The metric learning strategy is applied in conjunction with *k*-NN classification and novel sample selection criteria.

Specifically, consider a hyperspectral image $X = \{\mathbf{x}_i\}_{i=1}^{n}$, where $x_i = [x_{i,1}, \ldots, x_{i,d}]$ is a sample in the original high d-dimensional feature space and n is the total number of pixels in I. Define $\mathbf{x}'_i = [x'_{i,1}, \ldots x'_{i,r}]$ as the same sample in the reduced r-dimensional feature space obtained by adopting a feature extraction strategy. A training set $L = x_i, y_i{}_{i=1}^{l}$ is constructed by selecting l samples from X and assigning corresponding discrete labels y_i (where $y_i \in 1, \ldots, \Omega$, and Ω is the number of thematic classes). $U = x_i{}_{i=l+1}^{l+u}$ is defined as the set of u remaining labeled samples, that is, $U = X - L$ and $n = u + l$. The supervised dimensionality reduction strategy exploits the training set L to generate a low-dimensional feature space X' from X, and an AL strategy, where the most uncertain samples are selected and labeled, is then applied on X'. The process is iterated until a convergence criterion is satisfied.

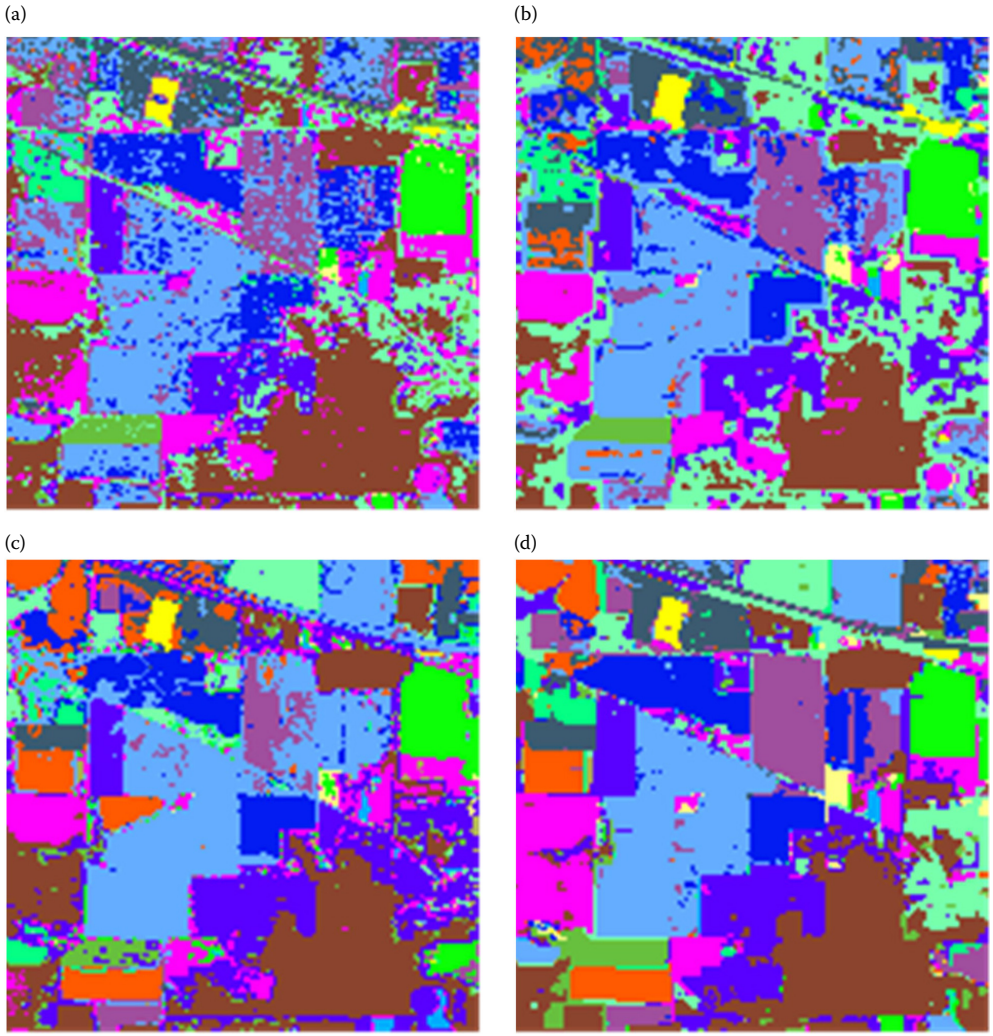


FIGURE 3.21 Classification maps achieved on the Indian Pine 1992 data set. (a) *Noseg*, (b) *Win_AddFeat*, (c) *EMP*, (d) *Adseg_AddFeat* + *AddSamp*. (Adapted from Z. Zhang et al. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 2, pp. 640–654, 2016. [98])

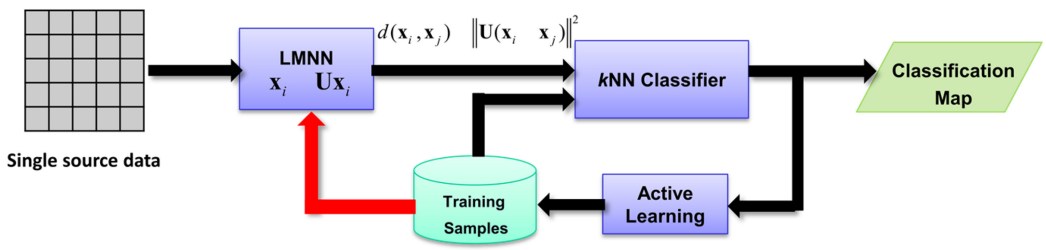


FIGURE 3.22 Flow chart illustrating the active-metric learning approach for supervised classification. (Adapted from E. Pasolli et al. *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 4, pp. 1925–1939, 2016. [122])

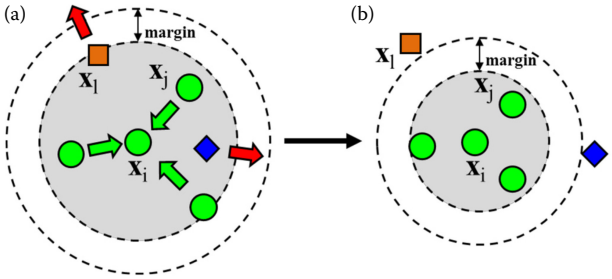


FIGURE 3.23 Schematic representation of LMNN metric learning method. (a) Original feature space. (b) Feature space after training. After training, *K* similar samples (in green) are separated from the dissimilar ones by a unit margin. Local neighborhood in gray. (Adapted from E. Pasolli et al. *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 4, pp. 1925–1939, 2016. [122])

The dimensionality reduction step associated with the LMNN algorithm [125], which is schematically represented in Figure 3.23, is implemented in conjunction with the Mahalanobis distance [126] and extended to improve classification performance, in terms of accuracy and computational time: (1) dimensionality reduction is incorporated directly into the objective function optimization process, (2) the optimization process is iterated through a multipass approach, and (3) *k*-NN search is accelerated through ball tree formulation.

The original method, which can only handle single input features (e.g., pure spectral features) and does not specifically accommodate multiple feature scenarios, is extended in [123], where multiple feature types are concatenated by extending LMNN to heterogeneous multimetric learning (HMML) [127]. The reduced feature space is obtained for each feature type by adopting a modified version of the HMML algorithm, and. AL is then applied in the resulting single feature space in conjunction with *k*-NN classification. Further improvements were obtained in [124] via a regularized multimetric AL frameworkto jointly learn distinct metrics for different feature types. The regularizer incorporates unlabeled data based on the neighborhood relationship, which also helps avoid overfitting at early AL stages. As the iterative process proceeds, the regularizer is updated through similarity propagation, thus taking advantage of informative labeled samples. Finally, multiple features are projected into a common feature space, in which a new batch-mode selection strategy that incorporates uncertainty and diversity criteria is used. Comparison on the Indian Pine 1992 data set among the different active-metric learning methods is reported in Figure 3.24.

3.6 SUMMARY AND FUTURE DIRECTIONS FOR CLASSIFICATION OF HYPERSPECTRAL IMAGES

This chapter has addressed key issues in classification of hyperspectral data and provided an overview of strategies to address these problems. As noted in the introduction, availability of hyperspectral imagery and associated ground reference data, including class labels, has been a significant hurdle to advancing classification methods focused on vegetation and agriculture croplands. This problem will be addressed in part by the upcoming launches of combinations of traditional near polar orbiting satellite missions and constellations of small satellites carrying hyperspectral cameras. The resulting improvement in temporal resolution of data will be particularly relevant to applications in agriculture and vegetation. Hyperspectral camera and global navigation satellite system (GNSS) technologies are also advancing, including miniaturization, making them viable sensors for UAVs and ground-based platforms that provide higher spatial resolution data that can be collected on demand at reduced cost. The resulting data sets will be enormous, motivating the need for advances in data processing and management, including on-board processing and analysis.

The high dimensionality of hyperspectral data, coupled with band redundancy, is a well-known obstacle for traditional parametric classifiers, particularly when the quantity of labeled data for

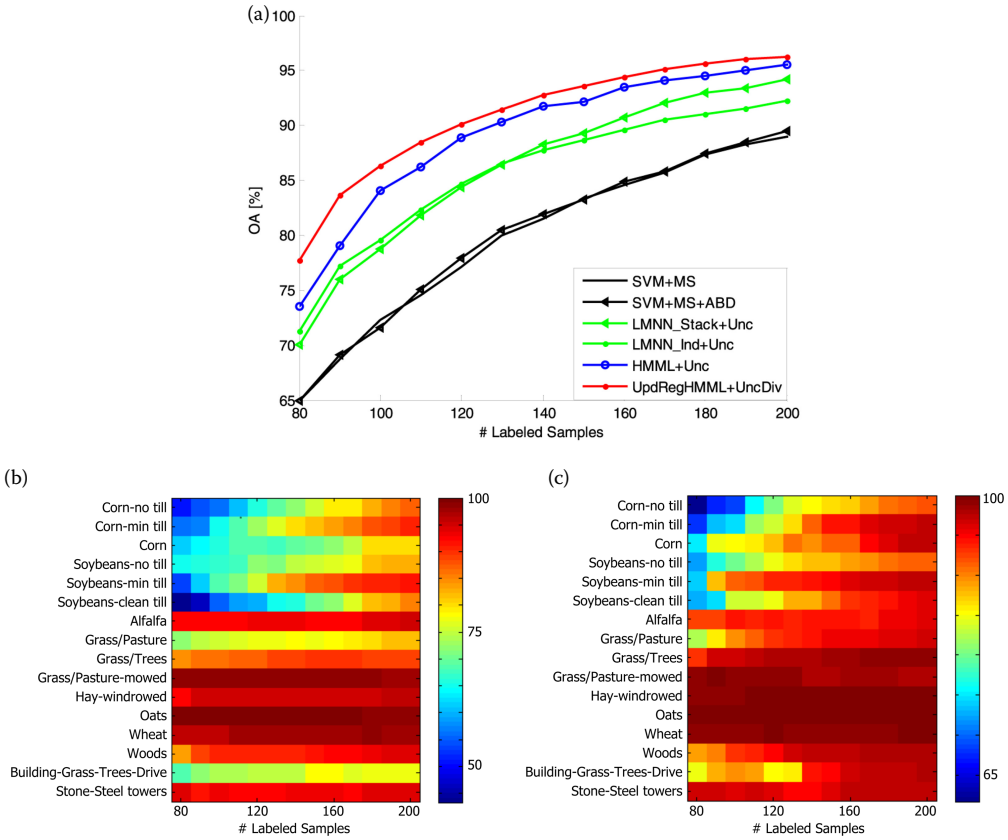


FIGURE 3.24 Comparisons among different active-metric learning methods on the Indian Pine 1992 data set. (a) Learning curve (overall accuracy in function of the number of labeled samples) for four strategies *UpdRegHMML+Unc* [124], *HMML+Unc* [123], *LMNN_Stack+Unc* [122], and *LMNN_Stack+Unc* [122] in addition to *SVM + MS* and *SVM + MS + ABD* [8]. Class-specific accuracies as a function of the number of labeled samples for (b) *SVM + MS* and (c) *UpdRegHMML + UncDiv*. (Adapted from Z. Zhang and M. M. Crawford, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 11, pp. 6594–6609, 2017. [124])

training is limited. Dimensionality reduction as a front-end processing step is both the most direct and most widely used approach to address this problem. Traditional methods that include direct feature selection as an optimization problem as either a supervised and unsupervised strategy (Section 3.3.1.1) and use of narrow band indices (Section 3.3.1.2) continue to be popular for classification, disease detection, and yield prediction in agriculture, as they are often correlated with plant vigor or specific chemistry-based responses at various stages of crop development. Both feature selection and vegetation indices have the advantage of being interpretable, but potentially ignore useful information in the rest of the spectrum. Alternatively, feature extraction approaches, which can exploit the full set of bands and inputs from other sources, are more popular for classification of single images, but suffer from interpretability and generalizability across multiple images. Traditional linear feature extraction methods (Section 3.3.1.3) are still widely used as inputs for classification and for other applications of hyperspectral data such as unmixing, in part because of their availability in commercial software. However, as discussed and illustrated in Section 3.3.1.4, nonlinear extraction approaches, particularly local graph-based methods, are promising for increasing class separation and thereby classification accuracy, although at increased computational cost.

Inclusion of spatial information can be extremely beneficial for classification of scenes containing natural vegetation or agriculture. Combining high spectral resolution data with spatial information

has made it possible to discriminate classes that are spectrally quite similar, even at the relatively coarse spatial resolution of AVIRIS and Hyperion, as the impact of within-class spatial variability is reduced. Higher spatial resolution data obtained by low-altitude aircraft, UAVs, and ground-based systems provide the capability to actually exploit higher-frequency spatial components in the data, which may be useful for improved class discrimination. Incorporation of texture measures as inputs (Section 3.3.2) and use of hierarchical segmentation approaches (Section 3.4.4) and new classification strategies such as deep learning (Section 3.4.3) all leverage multiscale spatial information for classification, as illustrated in Section 3.4.3 using the 2010 Indian Pine data. While no single strategy dominates as the best approach for inclusion of spatial information in a classification problem, example results in this chapter, as well as the literature, clearly demonstrate the importance of spatial context in classification of natural vegetation and agricultural images.

New models of inputs, as well as nonparametric data-driven approaches, are also receiving significant attention for classification of hyperspectral data. SVM classifiers (Section 3.4.1.1), which directly avoid the issue of non-Gaussian class conditional density functions and high-dimensional inputs, are now widely implemented, although proper parameter tuning is challenging and requires adequate quantities of training data. Multikernel extensions of traditional single-kernel SVMs (Section 3.5.1.1) are especially promising for multisource inputs, as demonstrated by the examples in this chapter and the literature. When coupled with feature extraction for dimensionality reduction, Gaussian mixture models implemented in a Bayesian framework can also provide an effective strategy for classifying hyperspectral data (Section 3.4.2). Deep learning approaches (Section 3.4.3), which are now being widely explored for classification of hyperspectral data, have the capability to exploit nonlinear relationships and interactions, but require very large data sets for training. The potential of deep learning for classification of hyperspectral data is really in its infancy in this domain. Early results have stimulated significant research for both algorithm development and applications, including approaches for leveraging the structure of hyperspectral spectra and new strategies for addressing the limited training data issue. The chapter also included examples of newly developed approaches for addressing challenges and new opportunities in hyperspectral data classification. Relative to training and application of classifiers, methods to tackle domain adaptation and transfer learning will be necessary as more hyperspectral data become available. We presented one strategy for feature alignment that yielded promising results, and referenced others (Section 3.5.1.2). We also included a novel example where training data acquired by a ground-based system were used to train a classifier that was applied over an extended area. It provided not only an opportunity for multiscale learning, but also a strategy for expanding the limited training data set.

Finally, we addressed the issue of limited training data through active and metric learning in Section 3.5.2. Active learning provides a flexible framework for initiating the classification process with a small number of training samples and augmenting the pool with unlabeled data. The framework can be implemented with front-end feature extraction from single or multiple sources and with appropriate backend classifiers. An example based on the Kennedy Space Center AVIRIS data illustrates the potential of the approach, including targeted learning related to classes that are difficult to discriminate (Section 3.5.2.1). One limitation of active learning is that the true labels of pixels identified for inclusion in the training set must be determined. One strategy for addressing this problem is illustrated using the original Indian Pine data, where active learning is coupled with hierarchical segmentation to leverage spatially homogeneous areas and semisupervised learning (Section 3.5.2.2). As shown in the quantitative and qualitative results, this is a particularly effective approach for the testbed data. Another potential problem of active learning is that features that are extracted from the initial small training set may not be optimal as learning progresses, necessitating updates that can be computationally intensive, particularly for nonlinear feature extractors. Metric learning (Section 3.5.2.3) provides a new strategy that naturally integrates updates to the reduced feature space into the active learning framework using the concept of the large-margin nearest neighbor principle, which naturally couples with a *k*-NN classifier (which can also be considered a limitation). Early results from active metric learning are promising for both single- and multiple-source input data.

While the current strategies for classification of hyperspectral data build on a rich foundation, significant advances are still needed. New classification algorithms whose data structures and architecture exploit hyperspectral data for multitemporal, multiscale studies of agriculture and vegetation are particularly needed to effectively utilize hyperspectral imagery, both as standalone methods and in conjunction with biophysical models. Classification methods have traditionally been developed in a stove-pipe approach by algorithm developers working in isolation from the application domain and with limited understanding of this domain. The next generation of classification systems must leverage the knowledge of collaborative, multidisciplinary research composed of both methodologically focused researchers and partners from the basic and applied earth and life sciences.

ACKNOWLEDGMENTS

The authors would like to acknowledge the graduate students and postdoctoral scholars at Purdue University and University of Houston for setting up the experiments and generating results with methods reviewed in this overview chapter. We would also like to thank Farideh Foroozandeh Shahraki at the University of Houston for her help with formatting the chapter.

REFERENCES

1. G. Hughes. "On the mean accuracy of statistical pattern recognizers," *IEEE Transactions on Information Theory*, vol. 14, no. 1, pp. 55–63, 1968.
2. L. David. "Hyperspectral image data analysis as a high dimensional signal processing problem," *IEEE Signal Processing Magazine*, vol. 19, no. 1, pp. 17–28, 2002.
3. S. Tadjudin and D. A. Landgrebe. "Covariance estimation with limited training samples," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, no. 4, pp. 2113–2118, 1999.
4. L. Breiman. "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
5. J. Ham, Y. Chen, M. M. Crawford, and J. Ghosh. "Investigation of the random forest framework for classification of hyperspectral data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 3, pp. 492–501, 2005.
6. X. Jia, B.-C. Kuo, and M. M. Crawford. "Feature mining for hyperspectral image classification," *Proceedings of the IEEE*, vol. 101, no. 3, pp. 676–697, 2013.
7. L. Bruzzone, M. Chi, and M. Marconcini. "Semisupervised support vector machines for classification of hyperspectral remote sensing images," in *Hyperspectral Data Exploitation: Theory and Applications*, New York, John Wiley & Sons, pp. 275–311, 2007.
8. D. Tuia, F. Ratle, F. Pacifici, M. Kanevski, and W. Emery. "Active learning methods for remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 7, pp. 2218–2232, 2009.
9. M. M. Crawford, D. Tuia, and H. L. Yang. "Active learning: Any value for classification of remotely sensed data?" *Proceedings of the IEEE*, vol. 101, no. 3, pp. 593–608, 2013.
10. H. Liu and L. Yu. "Toward integrating feature selection algorithms for classification and clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 4, pp. 491–502, 2005.
11. Q. Cheng, P. K. Varshney, and M. K. Arora. "Logistic regression for feature selection and soft classification of remote sensing data," *IEEE Geoscience and Remote Sensing Letters*, vol. 3, no. 4, pp. 491–494, 2006.
12. X. Chen, T. Fang, H. Huo, and D. Li. "Graph-based feature selection for object-oriented classification in VHR airborne imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 1, pp. 353–365, 2011.
13. X. Jia and J. A. Richards. "Segmented principal components transformation for efficient hyperspectral remote-sensing image display and classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, no. 1, pp. 538–542, 1999.
14. S. M. Davis, D. A. Landgrebe, T. L. Phillips, P. H. Swain, R. M. Hoffer, J. C. Lindenlaub, and L. F. Silva. *Remote Sensing: The Quantitative Approach*, New York, McGraw-Hill, 1978.
15. A. Paoli, F. Melgani, and E. Pasolli. "Clustering of hyperspectral images based on multiobjective particle swarm optimization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 12, pp. 4175–4188, 2009.

16. A. Ifarraguerri and M. W. Prairie. "Visual method for spectral band selection," *IEEE Geoscience and Remote Sensing Letters*, vol. 1, no. 2, pp. 101–106, 2004.
17. K. Fukunaga, *Introduction to Statistical Pattern Recognition*. Academic Press, 2013.
18. B. Paskaleva, M. M. Hayat, Z. Wang, J. S. Tyo, and S. Krishna. "Canonical correlation feature selection for sensors with overlapping bands: Theory and application," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 10, pp. 3346–3358, 2008.
19. T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley-Interscience, 2012.
20. Q. Du and H. Yang. "Similarity-based unsupervised band selection for hyperspectral image analysis," *IEEE Geoscience and Remote Sensing Letters*, vol. 5, no. 4, pp. 564–568, 2008.
21. P. Mitra, C. Murthy, and S. K. Pal. "Unsupervised feature selection using feature similarity," *IEEE Transactions on Pattern Analysis and Machine Ibranch and Boundntelligence*, vol. 24, no. 3, pp. 301–312, 2002.
22. J. M. Sotoca, F. Pla, and J. S. Sanchez. "Band selection in multispectral images by minimization of dependent information," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 37, no. 2, pp. 258–267, 2007.
23. L. Wang, X. Jia, and Y. Zhang. "A novel geometry-based feature-selection technique for hyperspectral imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 4, no. 1, pp. 171–175, 2007.
24. P. M. Narendra and K. Fukunaga. "A branch and bound algorithm for feature subset selection," *IEEE Transactions on Computers*, vol. 9, no. C-26, pp. 917–922, 1977.
25. P. Pudil, J. Novovičová, and J. Kittler. "Floating search methods in feature selection," *Pattern Recognition Letters*, vol. 15, no. 11, pp. 1119–1125, 1994.
26. H. Yao and L. Tian. "A genetic-algorithm-based selective principal component analysis (GA-SPCA) method for high-dimensional data feature extraction," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 41, no. 6, pp. 1469–1478, 2003.
27. A. Paoli, F. Melgani, and E. Pasolli. "Swarm intelligence for unsupervised classification of hyperspectral images," in *IEEE International Symposium on Geoscience and Remote Sensing*, vol. 5, 2009, pp. V–96.
28. L. Zhang, Y. Zhong, B. Huang, J. Gong, and P. Li. "Dimensionality reduction based on clonal selection for hyperspectral imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 12, pp. 4172–4186, 2007.
29. P. S. Thenkabail, I. Mariotto, M. K. Gumma, E. M. Middleton, D. R. Landis, and K. F. Huemmrich. "Selection of hyperspectral narrowbands (HNBS) and composition of hyperspectral two-band vegetation indices (HVIS) for biophysical characterization and discrimination of crop types using field reflectance and Hyperion/EO-1 data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 6, no. 2, pp. 427–439, 2013.
30. J. W. White, P. Andrade-Sanchez, M. A. Gore, K. F. Bronson, T. A. Coffelt, M. M. Conley, K. A. Feldmann, A. N. French, J. T. Heun, D. J. Hunsaker et al. "Field-based phenomics for plant genetics research," *Field Crops Research*, vol. 133, pp. 101–112, 2012.
31. S. Kumar, J. Ghosh, and M. M. Crawford. "Best-bases feature extraction algorithms for classification of hyperspectral data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 7, pp. 1368–1379, 2001.
32. S. De Backer, P. Kempeneers, W. Debruyn, and P. Scheunders. "A band selection technique for spectral classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 2, no. 3, pp. 319–323, 2005.
33. S. B. Serpico and G. Moser. "Extraction of spectral channels from hyperspectral images for classification purposes," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 2, pp. 484–495, 2007.
34. D. A. Landgrebe, *Signal Theory Methods in Multispectral Remote Sensing*. John Wiley & Sons, 2005.
35. B.-C. Kuo and D. A. Landgrebe. "Nonparametric weighted feature extraction for classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 5, pp. 1096–1105, 2004.
36. K. Fukunaga and J. Mantock. "Nonparametric discriminant analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 6, pp. 671–678, 1983.
37. H.-Y. Huang and B.-C. Kuo. "Double nearest proportion feature extraction for hyperspectral-image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 11, pp. 4034–4046, 2010.
38. I. T. Jolliffe. "Principal component analysis and factor analysis," in *Principal Component Analysis*. Springer, 1986, pp. 115–128.
39. H. Oja and K. Nordhausen. "Independent component analysis," *Encyclopedia of Environmetrics*, 2001.
40. L. O. Jimenez-Rodriguez, E. Arzuaga-Cruz, and M. Velez-Reyes. "Unsupervised linear feature-extraction methods and their effects in the classification of high-dimensional data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 2, pp. 469–483, 2007.

41. M. Cui and S. Prasad. "Angular discriminant analysis for hyperspectral image classification," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 6, pp. 1003–1015, 2015.
42. M. Cui and S. Prasad. "Spectral-angle-based discriminant analysis of hyperspectral data for robustness to varying illumination," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 9, pp. 4203–4214, 2016.
43. S. Mukherjee, M. Cui, and S. Prasad. "Spatially constrained semisupervised local angular discriminant analysis for hyperspectral images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 4, pp. 1203–1212, 2018.
44. C. M. Bachmann, T. L. Ainsworth, and R. A. Fusina. "Improved manifold coordinate representations of large-scale hyperspectral scenes," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 10, pp. 2786–2803, 2006.
45. D. Lunga, S. Prasad, M. Crawford, and O. Ersoy. "Manifold-learning-based feature extraction for classification of hyperspectral data: A review of advances in manifold learning," *IEEE Signal Processing Magazine*, vol. 31, no. 1, pp. 55–66, 2014.
46. J. B. Tenenbaum, V. De Silva, and J. C. Langford. "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.
47. B. Schölkopf, A. Smola, and K.-R. Müller. "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, no. 5, pp. 1299–1319, 1998.
48. S. T. Roweis and L. K. Saul. "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.
49. V. D. Silva and J. B. Tenenbaum. "Global versus local methods in nonlinear dimensionality reduction," in *Advances in Neural Information Processing Systems*, 2003, pp. 721–728.
50. L. K. Saul and S. T. Roweis. "Think globally, fit locally: Unsupervised learning of low dimensional manifolds," *Journal of Machine Learning Research*, vol. 4, pp. 119–155, 2003.
51. L. Ma, M. M. Crawford, X. Yang, and Y. Guo. "Local-manifold-learning-based graph construction for semisupervised hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 5, pp. 2832–2844, 2015.
52. H. L. Yang and M. M. Crawford. "Spectral and spatial proximity-based manifold alignment for multitemporal hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 1, pp. 51–64, 2016.
53. H. L. Yang and M. M. Crawford. "Domain adaptation with preservation of manifold geometry for hyperspectral image classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 2, pp. 543–555, 2016.
54. S. Yan, D. Xu, B. Zhang, and H.-J. Zhang. "Graph embedding: A general framework for dimensionality reduction," in *IEEE Computer Society Conference on CVPR*, vol. 2, 2005, pp. 830–837.
55. S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin. "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 40–51, 2007.
56. R. M. Haralick and K. Shanmugam. "Textural features for image classification," *IEEE Transactions on Systems, Man, and Cybernetics*, no. 6, pp. 610–621, 1973.
57. A. Baraldi and F. Parmiggiani. "An investigation of the textural characteristics associated with gray level cooccurrence matrix statistical parameters," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 33, no. 2, pp. 293–304, 1995.
58. J. Li, P. R. Marpu, A. Plaza, J. M. Bioucas-Dias, and J. A. Benediktsson. "Generalized composite kernel framework for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 9, pp. 4816–4829, 2013.
59. M. Dalla Mura, J. Atli Benediktsson, B. Waske, and L. Bruzzone. "Extended profiles with morphological attribute filters for the analysis of hyperspectral data," *International Journal of Remote Sensing*, vol. 31, no. 22, pp. 5975–5991, 2010.
60. D. Landgrebe. "Hyperspectral image data analysis," *IEEE Signal Processing Magazine*, vol. 19, no. 1, pp. 17–28, 2002.
61. G. Shaw and D. Manolakis. "Signal processing for hyperspectral image exploitation," *IEEE Signal Processing Magazine*, vol. 19, no. 1, pp. 12–16, 2002.
62. N. Keshava and J. Mustard. "Spectral unmixing," *IEEE Signal Processing Magazine*, vol. 19, no. 1, pp. 44–57, 2002.
63. D. Manolakis and G. Shaw. "Detection algorithms for hyperspectral imaging applications," *IEEE Signal Processing Magazine*, vol. 19, no. 1, pp. 29–43, 2002.

64. S. Prasad, H. Wu, and J. Fowler. "Compressive data fusion for multi-sensor image analysis," in *Proceedings of the IEEE International Conference on Image Processing*, October 2014, pp. 5032–5036.
65. W. Li, S. Prasad, E. W. Tramel, J. E. Fowler, and Q. Du. "Decision fusion for hyperspectral image classification based on minimum-distance classifiers in the wavelet domain," in *Proceedings of the Signal and Information Processing China Summit & International Conference*, 2014, pp. 162–165.
66. S. Prasad, M. Cui, W. Li, and J. Fowler. "Segmented mixture of Gaussian classification for robust sub-pixel hyperspectral ATR," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, no. 1, pp. 138–142, 2014.
67. S. Prasad, W. Li, J. E. Fowler, and L. M. Bruce. "Information fusion in the redundant-wavelet-transform domain for noise-robust hyperspectral classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 99, pp. 3474–3486, 2012.
68. W. Li, S. Prasad, and J. E. Fowler. "Classification and reconstruction from random projections for hyperspectral imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 2, pp. 1–11, 2012.
69. D. Tuia, G. Camps-Valls, G. Matasci, and M. Kanevski. "Learning relevant image features with multiple-kernel classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 10, pp. 3780–3791, 2010.
70. D. Tuia, F. Ratle, A. Pozdnoukhov, and G. Camps-Valls. "Multisource composite kernels for urban- image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 7, no. 1, pp. 88–92, January 2010.
71. M. Cui and S. Prasad. "Class dependent sparse representation classifier for robust hyperspectral image classification," *IEEE Transactions on Geosciences and Remote Sensing*, vol. 53, no. 5, pp. 2683–2695, May 2015.
72. M. Cui and S. Prasad. "Multiscale sparse representation classification for robust hyperspectral image analysis," in *Proceedings of the Global Conference on Signal and Information Processing*, 2013, pp. 969–972.
73. N. Segata, E. Pasolli, F. Melgani, and E. Blanzieri. "Local SVM approaches for fast and accurate classification of remote-sensing images," *International Journal of Remote Sensing*, vol. 33, no. 19, pp. 6186–6201, 2012.
74. G. Camps-Valls and L. Bruzzone. "Kernel-based methods for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 6, pp. 1351–1362, June 2004.
75. H. Akaike. "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.
76. C. E. Rasmussen. "The infinite Gaussian mixture model," in *Advances in Neural Information Processing Systems*, vol. 12, 1999, pp. 554–560.
77. H. Wu and S. Prasad. "Dirichlet process based active learning and discovery of unknown classes for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 8, pp. 4882–4895, 2016.
78. W. Li, S. Prasad, J. Fowler, and L. Bruce. "Locality-preserving dimensionality reduction and classification for hyperspectral image analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 4, pp. 1185–1198, 2012.
79. A. Krizhevsky, I. Sutskever, and G. E. Hinton. "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.
80. R. Girshick, J. Donahue, T. Darrell, and J. Malik. "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.
81. K. Simonyan and A. Zisserman. "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
82. A. Graves, A.-R. Mohamed, and G. Hinton. "Speech recognition with deep recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 6645–6649.
83. H. Sak, A. Senior, K. Rao, and F. Beaufays. "Fast and accurate recurrent neural network acoustic models for speech recognition," *arXiv preprint arXiv:1507.06947*, 2015.
84. P. J. Werbos. "Backpropagation through time: What it does and how to do it," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.
85. Z. Zuo, B. Shuai, G. Wang, X. Liu, X. Wang, B. Wang, and Y. Chen. "Convolutional recurrent neural networks: Learning spatial dependencies for image representation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 18–26.
86. H. Wu and S. Prasad. "Convolutional recurrent neural networks for hyperspectral data classification," *Remote Sensing*, vol. 9, no. 3, p. 298, 2017.

87. J. C. Tilton, S. Aksoy, and Y. Tarabalka. "Image segmentation algorithms for land categorization," in *Remotely Sensed Data Characterization, Classification, and Accuracies*, 2015, pp. 317–342.
88. F. A. Kruse, A. Lefkoff, J. Boardman, K. Heidebrecht, A. Shapiro, P. Barloon, and A. Goetz. "The spectral image processing system (SIPS), interactive visualization and analysis of imaging spectrometer data," *Remote Sensing of Environment*, vol. 44, no. 2–3, pp. 145–163, 1993.
89. J.-M. Beaulieu and M. Goldberg. "Hierarchy in picture segmentation: A stepwise optimization approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 2, pp. 150–163, 1989.
90. T. Kurita. "An efficient agglomerative clustering algorithm for region growing," in *IAPR Workshop on Mach. Vis. Appl. Citeseer*, 1994.
91. J. Williams. "Algorithm 232—Heapsort," *Communications of the ACM*, vol. 7, no. 6, pp. 347–348, 1964.
92. J. C. Tilton. "Image segmentation by region growing and spectral clustering with a natural convergence criterion," in *Proceedings of the Geoscience and Remote Sensing Symposium*, vol. 4, 1998, pp. 1766–1768.
93. J. C. Tilton, Y. Tarabalka, P. M. Montesano, and E. Gofman. "Best merge region-growing segmentation with integrated nonadjacent region object aggregation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 11, pp. 4454–4467, 2012.
94. J. C. Tilton, *Parallel Implementation of the Recursive Approximation of an Unsupervised Hierarchical Segmentation Algorithm*, New York, Chapman & Hall, 2007.
95. J. C. Tilton, E. B. de Colstoun, R. E. Wolfe, B. Tan, and C. Huang. "Generating ground reference data for a global impervious surface survey," in *IEEE International Geoscience and Remote Sensing Symposium*, 2012, pp. 5993–5996.
96. E. C. B. de Colstoun, C. Huang, P. Wang, J. C. Tilton, B. Tan, J. Phillips, S. Niemczura, P.-Y. Ling, and R. Wolfe. "Documentation for the global man-made impervious surface (GMIS) dataset from LANDSAT," 2017.
97. R. Massey, T. T. Sankey, K. Yadav, R. G. Congalton, and J. C. Tilton. "Integrating cloud-based workflows in continental-scale cropland extent classification," submitted to the *Remote Sensing of Environment*, 2017.
98. Z. Zhang, E. Pasolli, M. M. Crawford, and J. C. Tilton. "An active learning framework for hyperspectral image classification using hierarchical segmentation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 2, pp. 640–654, 2016.
99. A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet. "SimpleMKL," *Journal of Machine Learning Research*, vol. 9, no. November, pp. 2491–2521, 2008.
100. Y. Zhang, L. Yang, S. Prasad, E. Pasolli, J. Jung, and M. Crawford. "Ensemble multiple kernel active learning for classification of multisource remote sensing data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 2, 2015.
101. D. Tuia, J. Munoz-Mari, L. Gómez-Chova, and J. Malo. "Graph matching for adaptation in remote sensing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 1, pp. 329–341, 2013.
102. X. Zhou and S. Prasad. "Domain adaptation for robust classification of disparate hyperspectral images," *IEEE Transactions on Computational Imaging*, vol. 3, no. 4, pp. 822–836, December 2017.
103. G. Schohn and D. Cohn. "Less is more: Active learning with support vector machines," in *ICML. Citeseer*, 2000, pp. 839–846.
104. P. Mitra, B. Uma Shankar, and S. Pal. "Segmentation of multispectral remote sensing images using active support vector machines," *Pattern Recognition Letters*, vol. 25, no. 9, pp. 1067–1074, 2004.
105. L. Copa, D. Tuia, M. Volpi, and M. Kanevski. "Unbiased query-by-bagging active learning for VHR image classification," in *Image and Signal Processing for Remote Sensing XVI, vol. 7830. International Society for Optics and Photonics*, 2010, p. 78300K.
106. W. Di and M. M. Crawford. "View generation for multiview maximum disagreement based active learning for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 5, pp. 1942–1954, 2012.
107. T. Luo, K. Kramer, D. B. Goldgof, L. O. Hall, S. Samson, A. Remsen, and T. Hopkins. "Active learning to recognize multiple types of plankton," *Journal of Machine Learning Research*, vol. 6, no. April, pp. 589–613, 2005.
108. S. Rajan, J. Ghosh, and M. Crawford. "An active learning approach to hyperspectral data classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 4, pp. 1231–1242, 2008.
109. E. Pasolli, F. Melgani, D. Tuia, F. Pacifici, and W. J. Emery. "SVM active learning approach for image classification using spatial information," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 4, pp. 2217–2233, 2014.
110. Q. Shi, B. Du, and L. Zhang. "Spatial coherence-based batch-mode active learning for remote sensing image classification," *IEEE Transactions on Image Processing*, vol. 24, no. 7, pp. 2037–2050, 2015.

111. D. Tuia, E. Pasolli, and W. J. Emery. "Using active learning to adapt remote sensing image classifiers," *Remote Sensing of Environment*, vol. 115, no. 9, pp. 2232–2242, 2011.
112. E. Pasolli, F. Melgani, N. Alajlan, and N. Conci. "Optical image classification: A ground-truth design framework," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 6, pp. 3580–3597, 2013.
113. N. Alajlan, E. Pasolli, F. Melgani, and A. Franzoso. "Large-scale image classification using active learning," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, no. 1, pp. 259–263, 2014.
114. D. Tuia, M. Volpi, L. Copa, M. Kanevski, and J. Munoz-Mari. "A survey of active learning algorithms for supervised remote sensing image classification," *IEEE Journal of Selected Topics in Signal Processing*, no. 99, pp. 1–1, 2011.
115. J. Li, J. M. Bioucas-Dias, and A. Plaza. "Hyperspectral image segmentation using a new Bayesian approach with active learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 10, pp. 3947–3960, 2011.
116. J. Li, J. M. Bioucas-Dias, and A. Plaza. "Spectral–spatial classification of hyperspectral data using loopy belief propagation and active learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 2, pp. 844–856, 2013.
117. A. Stumpf, N. Lachiche, J.-P. Malet, N. Kerle, and A. Puissant. "Active learning in the spatial domain for remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 5, pp. 2492–2507, 2014.
118. P. Salembier and L. Garrido. "Binary partition tree as an efficient representation for image processing, segmentation, and information retrieval," *IEEE Transactions on Image Processing*, vol. 9, no. 4, pp. 561–576, 2000.
119. S. Valero, P. Salembier, and J. Chanussot. "Hyperspectral image representation and processing with binary partition trees," *IEEE Transactions on Image Processing*, vol. 22, no. 4, pp. 1430–1443, 2013.
120. J. Munoz-Mari, D. Tuia, and G. Camps-Valls. "Semisupervised classification of remote sensing images with active queries," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 10, pp. 3751–3763, 2012.
121. J. Jung, E. Pasolli, S. Prasad, J. C. Tilton, and M. M. Crawford. "A framework for land cover classification using discrete return LiDAR data: Adopting pseudo-waveform and hierarchical segmentation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 2, pp. 491–502, 2014.
122. E. Pasolli, H. L. Yang, and M. M. Crawford. "Active-metric learning for classification of remotely sensed hyperspectral images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 4, pp. 1925–1939, 2016.
123. Z. Zhang, E. Pasolli, H. L. Yang, and M. M. Crawford. "Multimetric active learning for classification of remote sensing data," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 7, pp. 1007–1011, 2016.
124. Z. Zhang and M. M. Crawford. "A batch-mode regularized multimetric active learning framework for classification of hyperspectral images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 11, pp. 6594–6609, 2017.
125. K. Q. Weinberger and L. K. Saul. "Distance metric learning for large margin nearest neighbor classification," *Journal of Machine Learning Research*, vol. 10, no. February, pp. 207–244, 2009.
126. A. Bellet, A. Habrard, and M. Sebban. "A survey on metric learning for feature vectors and structured data," *arXiv preprint arXiv:1306.6709*, 2013.
127. H. Zhang, T. S. Huang, N. M. Nasrabadi, and Y. Zhang. "Heterogeneous multi-metric learning for multi-sensor fusion," in *Proceedings of the 14th International Conference on Information Fusion*, IEEE, 2011, pp. 1–8.