On the influence of feature reduction for the classification of hyperspectral images based on the extended morphological profile

Thibaut Castaings, Bjorn Waske, Jon Atli Benediktsson, Jocelyn Chanussot

International Journal of Remote Sensing
Vol. 31, No. 22, 20 November 2010, 5921–5939

Taylor & Francis
Taylor & Francis Group

On the influence of feature reduction for the classification of hyperspectral images based on the extended morphological profile

THIBAUT CASTAINGS†‡, BJÖRN WASKE§, JÓN ATLI BENEDIKTSSON*† and JOCELYN CHANUSSOT‡

†Faculty of Electrical and Computer Engineering, University of Iceland, 107 Reykjavik, Iceland
‡GIPSA-Lab, Grenoble Institute of Technology, 38402 Saint Martin d'Hères, France
§Institute of Geodesy and Geoinformation, Faculty of Agriculture, University of Bonn, 53115 Bonn, Germany

In this study we investigated the classification of hyperspectral data with high spatial resolution. Previously, methods that generate a so-called extended morphological profile (EMP) from the principal components of an image have been proposed to create base images for morphological transformations. However, it can be assumed that the feature reduction (FR) may have a significant effect on the accuracy of the classification of the EMP. We therefore investigated the effect of different FR methods on the generation and classification of the EMP of hyperspectral images from urban areas, using a machine learning-based algorithm for classification. The applied FR methods include: principal component analysis (PCA), nonparametric weighted feature extraction (NWFE), decision boundary feature extraction (DBFE), Gaussian kernel PCA (KPCA) and Bhattacharyya distance feature selection (BDFS). Experiments were run with two classification algorithms: the support vector machine (SVM) and random forest (RF) algorithms. We demonstrate that the commonly used PCA approach seems to be nonoptimal in a large number of cases in terms of classification accuracy, and the other FR methods may be more suitable as preprocessing approaches for the EMP.

1. Introduction

High-resolution remotely sensed images from urban areas have recently become available. The classification of such images is challenging because urban areas often comprise a large number of different surface materials, and consequently the heterogeneity of urban images is relatively high. Moreover, different information classes can be made up of spectrally similar surface materials (Herold *et al.* 2004). In many studies, the classification accuracy can be improved by including spatial information in the image analysis. Several approaches have been introduced to extract spatial information from images, using texture measures, segment-based analysis or morphological operators.

In the context of image processing by mathematical morphology (MM), two fundamental operators, erosion and dilation, are applied to grey-scale images with a set of known shapes, the so-called structuring elements (SEs). Erosion replaces each element in a greyscale image by the minimal element in the set of neighbours specified by the nonzero elements in an SE, whereas dilation replaces each element in a

*Corresponding author. Email: benedikt@hi.is

International Journal of Remote Sensing
ISSN 0143-1161 print/ISSN 1366-5901 online
http://www.tandf.co.uk/journals
DOI: 10.1080/01431161.2010.512313

greyscale image by the maximal element in the set of neighbours specified by the nonzero elements in an SE. Erosion and dilation define most of the other morphological operators. Well-known morphological operators are 'opening', which is the dilation of an eroded image, and 'closing', which erodes a dilated image. The morphological opening removes the locally bright objects in the image, whereas the morphological closing removes objects that are darker than the surrounding area. The use of geodesic reconstruction during an opening or closing operation prevents the appearance of shape noise. MM was initiated by the works of Matheron (1967, 1975) and Serra (1982, 1988). Example of more recent studies on hyperspectral images are those by Soille and Pesaresi (2002) and Soille (2003).

The morphological operators require the definition of an SE with a fixed size. However, the *a priori* definition of the size is often problematic and using only one size is often inefficient because of the various object sizes within the image. One size of the SE might be more applicable to detecting objects from a specific information class whereas another size might be better for describing another class. Thus, the use of multiscale approaches that relies on different SE sizes seems to be suitable (Soille 2003). The use of a series of SEs with different sizes is called granulometry (Serra 1982). Granulometry by opening is based on the sequential use of opening operations with an SE of increasing size; as the result the image is successively simplified.

Pesaresi and Benediktsson (2001) successfully used a composition of opening and closing operations to build a so-called morphological profile (MP), as the input of a neural network for the classification of panchromatic data (a single-band image). Benediktsson *et al.* (2003) defined image features by their morphological intrinsic characteristics, instead of using their boundary as in Pesaresi and Benediktsson (2001). In doing so, the classification accuracy of the panchromatic high-resolution images was significantly increased.

MM exhibits further modification and is used for different applications. Akcay and Aksoy (2008), for example, performed automatic object detection in high-resolution images by combining spectral information with structural information by a segmentation method, which is based on morphological operations with SEs of increasing sizes. Bellens *et al.* (2008) used directional MPs for an improved classification of very high-resolution urban imagery. Tuia *et al.* (2009) investigated the relevance of morphological operators for the classification of land use in urban scenes, using support vector machines (SVMs) and a feature selection (FS) algorithm. Different morphological operators were used, and the type and scale of the operators are discussed.

High-resolution hyperspectral remote sensing data from urban areas have recently become available, and consequently the number of applications that are based on hyperspectral data have increased significantly over the past years. Such data provide detailed structural and spectral information about urban scenes. However, the use of each band of a hyperspectral data set for the generation of an MP does not seem to be feasible with regard to the high dimensionality as well as the redundancy within the bands. Thus, the MP method had to be extended to these types of data. To use both spatial and spectral information in classification, Benediktsson *et al.* (2005) proposed an approach based on an extended morphological profile (EMP). An EMP is a series of MPs built from each of the most significant principal components resulting from a principal component analysis (PCA). The use of the PCA aims to avoid the generation of a very high-dimensional data set and a significant increase in redundancy.

As the construction of an MP is based on a range of increasing size of morphological operators, the resulting data sets can be high dimensional and may contain

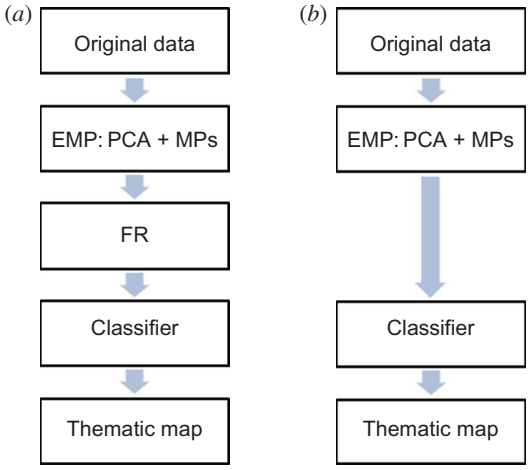


Figure 1. Two standard approaches: (*a*) a feature reduction (FR) method can be applied on the output of the extended morphological profile (EMP) or (*b*) the output of the EMP can be used directly as input for the classifier.

redundant and irrelevant information. Therefore, effective classification schemes consist of performing an additional feature extraction (FE), after the generation of the EMP, and in using classifiers that can handle such data sets. This is illustrated in figure 1. The use of different additional FE methods that have been applied on the EMP has been discussed previously (e.g. Benediktsson *et al.* 2003, Fauvel *et al.* 2008). The use of common statistical classifiers is often limited in this context because of the high dimensionality and complexity of the spectral–spatial data set. Thus, the use of more advanced approaches, such as neural networks (Benediktsson *et al.* 2005), SVM (Fauvel *et al.* 2008) and random forest (RF) classifiers (Gislason *et al.* 2006) is more suitable in this context.

Although several studies have discussed the effect of FE after generating the EMP, the use of different FE methods for the selection of the input data for the EMP has not yet been investigated. The PCA FE method is theoretically optimal, in the mean squared sense, for the representation of data and it enables a good visualization of the resulting data. However, it may be limited in the context of image classification. Thus, other feature reduction (FR) methods may be more suitable for the EMP to improve the overall classification accuracy.

In this study, we investigate the effect of different FR methods on the generation and classification of the EMP of hyperspectral images from urban areas. In addition to PCA, we consider the nonparametric weighted FE (NWFE), decision boundary FE (DBFE), Gaussian kernel PCA (KPCA) FE and the Bhattacharyya distance feature selection (BDFS) methods.

2. EMP

The aim of building an MP (Pesaresi and Benediktsson 2001) is to obtain information about the objects contained in an image, x. For this purpose, several opening and closing operations need to be applied to the input image SE of increasing size (see figure 2). Let $(\mathrm{OP})_n(x)$ (resp. $(\mathrm{CL})_n(x)$) be the image resulting from an opening (resp.

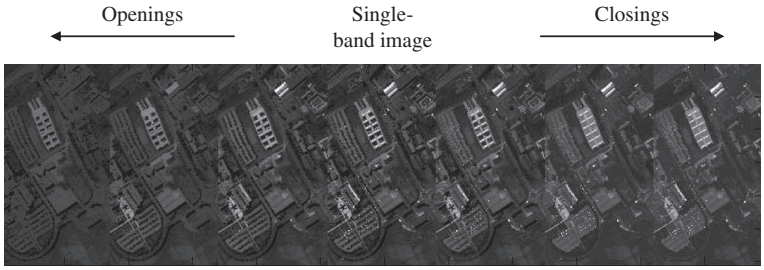


Figure 2. Example of a morphological profile (MP).

closing), with a disc-shaped SE of radius n on the single-band image x. This enables the size of the structures in the image to be taken into account; the image resulting from a closing operation will remove the locally dark structures, larger than $2n$ pixels. As a consequence, several opening and closing operations are needed to discriminate the structures according to their sizes and brightness.

Thus, let the MP of x be defined as:

$$\mathrm{MP}(x) = \left\{(\mathrm{CL})_m(x), \ldots, x, \ldots, (\mathrm{OP})_m(x)\right\} \tag{1}$$

where m is a fixed, arbitrary chosen integer.

As discussed in Benediktsson *et al.* (2005), the MP approach of Pesaresi and Benediktsson (2001) can be extended for hyperspectral data. However, this is a challenging task, as the proper definition of morphological operators in very high dimensional spaces is not straightforward. The proposed strategy consists in defining the EMP of X as the concatenation of the different $\mathrm{MP}(x_i)$:

$$\mathrm{EMP}(X) = \{(\mathrm{MP})(x_i), i \in [[1, N]]\} \tag{2}$$

At this point, the EMP contains both spectral and spatial information. Nevertheless, building an EMP given a particular set of raw data, such as an airborne multiband image, which could have more than 100 bands, would result in $(2m + 1) \times 100$-band data. To avoid this, it is common to use the principal components (resulting from a PCA) as input of the EMP. The PCA reduces the redundancy and, as a result, reduces the dimension of the raw data, keeping most of the spectral information. Most of the information remains in the first few principal components, and the resulting EMP may contain a small enough number of bands to enable a reasonable computational load and dimension of the remaining data.

In addition to the PCA, many other FR methods have been developed. They are reviewed in the following section.

3. FR methods

Several FR methods have been developed to reduce the number of features of the data to be classified while keeping the discriminating information. Two families of FR methods have been proposed. Whereas feature selection (FS), such as the BDFS, refers to a selection of a subset of relevant features, FE methods (e.g. PCA, NWFE, DBFE, KPCA) combine and transform the original feature space to obtain a relevant representation of the data in a lower-dimensional space. Thus, the original feature

space is preserved by FS (i.e. a specific band still corresponds to a specific wavelength) and is generally modified by the latter approach. A feature selection method aims at selecting the best bands, in which the maximum of the separability remains, that is:

$$Y = \left\{ X^{(i)}, i \in [[1, N]] \right\} \tag{3}$$

where Y are the output data, $X^{(i)}$ the input bands and N a user-defined integer. By contrast, a linear FE method transforms the data by computing a set of combinations of bands from the input data (new features space):

$$Y = \mathbf{\Phi}^{\mathrm{T}} X \tag{4}$$

where X is the input data, Y the output data, $\mathbf{\Phi}$ the transformation matrix with the feature vectors and T the transpose operator.

In the following sections we explain the general principles of different FR methods, namely the PCA, KPCA, NWFE, DBFE and BDFS. They are discussed in more detail in Landgrebe (2002).

3.1 *PCA FE*

The PCA is probably the most commonly used FE method in the field of remote sensing. It has been shown to be optimal for representation in terms of the mean squared error. However, it is not optimal for classification because it does not take class-specific information into account. In contrast to the other FR methods presented in this study (except KPCA), the PCA deals with the correlation between variables only, which may not necessarily improve the interclass separability. The principle of PCA can be described as follows: from a set of vectors, it aims at finding the most stretching space, defined according to the minimization of correlation between the variables, onto which the input vectors will be projected. In other words, the input vector set is to be projected onto a space that maximizes its variance.

Let $\{p_i \in R^n \quad \text{with} \quad i \in [[1, \ldots, l]]\}$ be a set of centred pixels from an n-band image, containing at least l pixels, with R being the set of real numbers. The PCA solves the eigenvalue problem:

$$\lambda \boldsymbol{v} = [p_1\ p_2\ \ldots\ p_l] \cdot [p_1\ p_2\ \ldots\ p_l]^{\mathrm{T}} \boldsymbol{v}, \quad \text{subject to } \|\boldsymbol{v}\|_2 = 1 \tag{5}$$

where λ is an eigenvalue (scalar) and $\boldsymbol{v}$ is a corresponding $n \times 1$ eigenvector. Let all the obtained eigenvalues, λ, be sorted from the highest to the lowest, and pick out the corresponding eigenvectors. A projection onto the m first components is performed as:

$$\mathbf{x_{pc}} = [\boldsymbol{v}_1\ \boldsymbol{v}_2\ \ldots\ \boldsymbol{v}_m]^{\mathrm{T}} \cdot [p_1\ p_2\ \ldots\ p_l] \tag{6}$$

In the case of hyperspectral data, m is chosen such that the cumulative value (variance) of the m largest λs is some prespecified percentage of the cumulative value of all the λs, typically 95–99%. For a 100-band remote sensing image, m is often around 4 if 99% of the cumulative variance is used.

3.2 *Gaussian KPCA FE*

The principle of the KPCA is similar to the PCA but it aims at capturing higher-order statistics, mapping the input space into another space H.

$$\begin{aligned} \mathbf{\Phi} : R^n &\rightarrow H \\ x &\rightarrow \mathbf{\Phi}(x) \end{aligned}$$

The kernel PCA solves the following eigenvalue problem:

$$\lambda \boldsymbol{v} = \mathbf{K}\boldsymbol{v}, \quad \text{subject to } \|\boldsymbol{v}\|_2 = \frac{1}{\lambda} \tag{7}$$

where

$$\mathbf{K} = \begin{pmatrix} k(p_1,p_1) & k(p_1,p_2) & \dots & k(p_1,p_l) \\ k(p_2,p_1) & k(p_2,p_2) & \dots & \dots \\ \dots & \dots & \dots & \dots \\ k(p_l,p_1) & \dots & \dots & k(p_l,p_l) \end{pmatrix}$$

and k is the core, or kernel, of the KPCA (k may be a polynomial kernel, a Gaussian kernel, or any semidefinite function that introduces nonlinearity into the processing). As the classes in hyperspectral remote sensing data are often well approximated by Gaussian distributions (Richards and Jia 1999), the Gaussian kernel has been chosen for the study reported here:

$$\begin{aligned} k: \quad R^{2n} &\rightarrow R \\ \boldsymbol{x}, \boldsymbol{y} &\rightarrow \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{2\sigma^2}\right), \quad \text{with } \sigma \in R \end{aligned} \tag{8}$$

The kernel k in (8) computes a scalar from two input vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ belonging to R^n and uses a preset scalar parameter σ in the computation. The choice of a suitable σ value to process the KPCA is important. It has to be chosen so that the number of features containing a specified value (e.g. 99%) of the total cumulative variance in the resulting reduced feature data is about the same as the number of classes of interest, or at least of the same order, as stated in Fauvel *et al.* (2009). The projection onto the new space is similar to that of the PCA. For further details, refer to Fauvel *et al.* (2009).

3.3 *DBFE*

The DBFE was proposed by Lee and Landgrebe (1993, 1997) to find the best features of a given dimensionality and the best dimensionality to use. Let X be a set of vectors belonging to two classes. A decision boundary feature matrix ($\mathbf{\Sigma_{DB}}$) is defined from the DB based on:

$$\sum\nolimits_{\mathbf{DB}} = \frac{1}{L}\sum \boldsymbol{N}_i \boldsymbol{N}_i^{\mathrm{T}} \tag{9}$$

where L is the number of training samples and the $\boldsymbol{N}_i$ are vectors normal to the lines connecting each pair of training samples belonging to different classes. If the classification problem involves more than two classes, a DB feature matrix has to be calculated for each pair of classes, then to be averaged with the DB matrices obtained from other pairs of classes. The eigenvectors of the resulting averaged DB feature matrix are the new feature vectors obtained with the DBFE.

3.4 *NWFE*

Kuo and Landgrebe (2004) showed that the NWFE is a powerful FE method for pattern recognition, compared to discriminant analysis feature extraction (DAFE),

nonparametric discriminant analysis (NDA) and DA using Malina's criterion. It can be described as follows: let $x_k^{(i)}$ be a set of $\boldsymbol{k}$ vectors belonging to class i. The principle of the NWFE is to take into account the vectors themselves along with the means of the (weighted) vectors belonging to each class. Nonparametric between-class ($\mathbf{S_b}$) and within-class ($\mathbf{S_w}$) scatter matrices are defined as follows:

$$\mathbf{S_b} = \sum_{i=1}^{nc} \frac{P_i}{nc-1} \sum_{\substack{j=1 \\ j\neq 1}}^{nc} \sum_{k=1}^{n_i} \lambda_k^{(i,j)} \left(x_k^{(i)} - M_j\left(x_k^{(i)}\right)\right)\left(x_k^{(i)} - M_j\left(x_k^{(i)}\right)\right)^{\mathrm{T}} \tag{10}$$

$$\mathbf{S_w} = \sum_{i=1}^{nc} P_i \sum_{k=1}^{n_i} \lambda_k^{(i,j)} \left(x_k^{(i)} - M_i\left(x_k^{(i)}\right)\right)\left(x_k^{(i)} - M_i\left(x_k^{(i)}\right)\right)^{\mathrm{T}} \tag{11}$$

where P_i and n_i are the prior probabilities and the number of samples belonging to class i, and nc is the total number of classes. The $x_k^{(i)}$ refer to the kth sample from class i. Furthermore:

$$\lambda_k^{(i,j)} = \frac{\operatorname{dist}\left(x_k^{(i)}, M_j\left(x_k^{(i)}\right)\right)^{-1}}{\sum_{l=1}^{n_j} \operatorname{dist}\left(x_l^{(i)}, M_j\left(x_l^{(i)}\right)\right)^{-1}}, M_j\left(x_k^{(i)}\right) = \sum_{l=1}^{n_j} w_l^{(i,j)} x_l^{(i)} x_l^{(j)},$$
$$\text{where} \quad w_l^{(i,j)} = \frac{\operatorname{dist}\left(x_k^{(i)}, x_l^{(j)}\right)^{-1}}{\sum_{l=1}^{n_j} \operatorname{dist}\left(x_k^{(i)}, x_l^{(j)}\right)^{-1}} \tag{12}$$

The transformation matrix is obtained from the m eigenvectors of $\mathbf{S_w}^{-1}\mathbf{S_b}$ that correspond to the m largest eigenvalues.

3.5 *BDFS*

The principle of the BDFS (Landgrebe 2002) is to select a subset of bands from the raw data according to the interclass BD between these bands, given a training sample set. For this purpose, the BDs between each pair of classes have to be computed for each combination of a given number of bands from the raw data. Afterwards, the BDs may be averaged for each subset, and the best subset can be selected according to this criterion or to some other criterion (e.g. the minimum BD). Under the assumption of Gaussian distribution and for a given subset of bands from the raw data, the BD, B_{ab}, between class a and class b is computed by:

$$B_{ab} = \frac{1}{8}[\boldsymbol{\mu}_a - \boldsymbol{\mu}_b]^{\mathrm{T}} \left[\frac{\boldsymbol{\Sigma}_a + \boldsymbol{\Sigma}_b}{2}\right]^{-1} [\boldsymbol{\mu}_a - \boldsymbol{\mu}_b] + \frac{1}{2}\ln \frac{\left|\frac{1}{2}[\boldsymbol{\Sigma}_a + \boldsymbol{\Sigma}_b]\right|}{\sqrt{|\boldsymbol{\Sigma}_a||\boldsymbol{\Sigma}_b|}} \tag{13}$$

where $\boldsymbol{\mu}_a$ and $\boldsymbol{\mu}_b$ are the mean vectors for the respective classes, and $\boldsymbol{\Sigma}_a$ and $\boldsymbol{\Sigma}_b$ the covariance matrices.

4. Classification methods

4.1 *RF*

In several studies the multiple classifier system (MCS) RF has been used for the classification of hyperspectral data (Ham *et al.* 2005, Joelsson *et al.* 2005, Waske *et al.* 2009). An MCS is a combination of different classifier algorithms and a combination of variants of the same classifier. By training a classifier algorithm on modified input data (i.e. training samples or input features), different independent classifier outputs are generated. Afterwards, the individual outputs are combined by a voting scheme and often a majority vote is used. The basic idea of an MCS is based on the assumption that independent classifiers produce individual errors that are not produced by the majority of the other classifiers. Thus, a combination of these outputs can increase the classification accuracy (Polikar 2006).

Breiman's RF approach (Breiman 2001) is based on the combination of different decision-tree classifiers (Breiman *et al.* 1984). Each tree within the ensemble is trained only on a subset of the original training samples. Moreover, only a randomly selected subset of k features of the input data is used at each node of the tree (Breiman 2001). For classification of image data, each tree classifies a pixel and a majority vote is performed to determine the final class membership. The influence of k on classification accuracy is discussed in Bernard *et al.* (2009).

4.2 *SVM*

SVMs perform well in different remote sensing studies, even when high-dimensional imagery is classified with small training sample sets. In the context of remote sensing and MM, SVMs were used, for example, by Tuia *et al.* (2009) and Fauvel *et al.* (2008). Moreover, an SVM still exhibits further improvement and modification such as the extension to semisupervised approaches (Chi and Bruzzone 2007), for example with a focus on mislabelled training data (Bruzzone and Persello 2009) and limited number of training samples (Ghoggali *et al.* 2009) as well as the adaptation to spectral–spatial and multisource classification problems (Camps-Valls *et al.* 2006, 2008).

In the context of remote sensing the SVM approach performs better than, or at least as well as, other classifiers even when classifying high-dimensional data sets (Pal and Mather 2006). The general principle of the SVM is finding a separating hyperplane in a feature space induced by a kernel function (Vapnik 1998), such as a Gaussian kernel (the standard deviation γ is a user-defined parameter) or any positive semidefinite function. The decision function is found by solving a convex optimization problem, subject to a user-defined parameter C. The kernel trick (Vapnik 1998) enables the computations to be made in the original space. An introduction to SVM in the context of remote sensing is given in Huang *et al.* (2002), a detailed general introduction can be found in Burges (1998) and Cristianini and Shawe-Taylor (2000).

5. Experimental results

5.1 *Data sets*

Experiments were run on two data sets, namely the 'University Area' and 'Pavia Centre' from urban areas in the city of Pavia, Italy (figure 3). The University Area and the Pavia Centre data were obtained from the airborne optical Reflective Optics Systems Imaging Spectrometer (ROSIS)-03 sensor. The ROSIS-03 provides 115

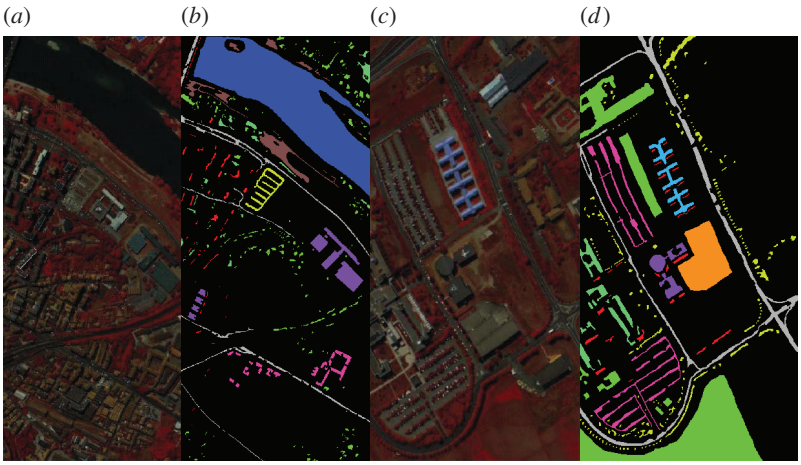


Figure 3. (*a*) Centre of Pavia data and (*b*) its ground truth map; (*c*) University of Pavia data and (*d*) its ground truth map.

Table 1. Classes and number of test samples of the Centre and University data sets.

Centre of Pavia			University of Pavia		
Class	Colour	Test samples	Class	Colour	Test samples
Water		65 278	Trees		3064
Trees		6503	Asphalt		6631
Asphalt		7493	Bitumen		1330
Bricks		2140	Gravel		2099
Bitumen		7287	Metal sheets		1345
Tiles		3122	Shadow		947
Shadow		2165	Bricks		3682
Meadows		2880	Meadows		18 649
Bare soil		6549	Bare soil		5029

bands with a nominative spectral coverage ranging from 0.43 to 0.86 μm. Its spatial resolution is 1.3 m per pixel.

The University Area data are 610 $\times$ 340 pixels and have 103 bands (12 bands were removed due to noise). Nine classes of interest are considered: Tree, Asphalt, Bitumen, Gravel, Metal sheet, Shadow, Brick, Meadow, and Bare soil (see table 1). The University Area is a low-density urban area and the data contain a large variety of shapes. The Pavia Centre data are 1096 $\times$ 489 pixels and have 102 bands (13 bands were removed due to noise). Nine classes of interest are considered: Water, Tree, Meadow, Brick, Soil, Asphalt, Bitumen, Tile, and Shadow. The Pavia Centre is a high-density urban area.

5.2 *Experimental setup*

The experiments were run using an SVM and an RF classifier. The classification with an SVM was performed with a MATLAB interface, which uses LIBSVM (available at

www.csie.ntu.edu.tw/~cjlin/libsvm). The classification with an RF was performed with a freely available MATLAB code available at www.pudn.com.

Because SVMs are defined as binary classifiers, a multiclass strategy is necessary to solve multiclass problems. In this study, the widely used one-against-one strategy was used, which generates a set of classifiers, one for each possible pair of classes. The parameters of C and γ were determined by a grid search, using cross-validation. Possible combinations were tested in a user-defined range and the best combinations were selected for the final classification.

The main parameters for the classification process with an RF are the number of features used at each node and the number of individual decision trees within the ensemble. After preliminary experiments, k was defined as 10, and 100 individual trees were used within the classifier system.

The experiments were run using the PCA, KPCA, NWFE, DBFE and BDFS. To investigate the impact of the individual FR methods in more detail, different experiments were conducted:

(1) Training sample sets with different numbers of samples were used: 25, 50, 100 and 200 samples per information class, respectively. These training sets were generated randomly from the ground truth data.
(2) Different numbers of selected features were tested. Each training sample set was used for the classification of different features sets, which were selected by different FR methods and contain different numbers of features, as presented in table 2.
(3) Two different classifiers were used: SVM and RF.

For the PCA and the DBFE, the number of bands was chosen to reflect over 99, 95, 90 and 80% of the cumulative variance of the data, respectively. The number of bands to selected from the NWFE was the same as the number of bands selected from the DBFE (so as to compare the effects of these two FE approaches). However, the cumulative variance reflected by the chosen number of bands selected from the NWFE was lower than that observed for the DBFE (up to 5 points lower). In table 2, nb1, nb2, nb3 and nb4 refer to the number of bands used for classification (e.g. nb1 = 4, nb3 = 3, nb2 = 2 and nb1 = 2 for PCA and the University of Pavia data).

After the PCA transformation, most of the variance (99, 98 and 90%) remained in the fourth, third or second principal components. In contrast to this, up to nine principal components were required by the KPCA to retain 99% of the variance, depending on the number of information classes (Fauvel *et al.* 2009). In the case of the BDFS, nb1 was chosen to be the same as for DBFE and NWFE (reflecting an average class separability of about 2). Then, nb2, nb3 and nb4 were chosen such that the separability decreased by about 0.5 from one specific number of bands to another. The number of features used with the BDFS were also chosen to be close to the number of features used with the DBFE and the NWFE.

Table 2. Number of bands used for the experiments: nb1, nb2, nb3 and nb4.

	PCA	KPCA	DBFE and NWFE	BDFS
University of Pavia	4, 3, 2, 2	9, 7, 5, 3	25, 16, 10, 7	25, 20, 15, 10
Centre of Pavia	3, 2, 2, 2	9, 7, 5, 3	20, 11, 8, 5	20, 15, 10, 5

Each classification (i.e. a specific training sample set and feature subset) was performed with two algorithms: the SVM and the RF. The overall accuracy (OA; the ratio of the total number of well-classified samples to the total number of samples), the average accuracy (AA; the mean of the class accuracies) and the class accuracy (the ratio of the number of well-classified samples of a class to the total number of samples of the same class) were determined, using an independent test set, which was identical for all classifications. To enhance the reliability of the experiments, all experiments were performed five times, using different, randomly selected training samples sets. Finally, the accuracy measures (OA, AA, class accuracy) were averaged.

5.3 *Results*

5.3.1 Classifications with the RF. In the case of the RF classification of the Centre of Pavia data (figure 4), the highest accuracies were achieved with the NWFE. Compared to the results achieved with 25 training samples per class for the PCA, the accuracy increased from 96.5% to 97.5% using the NWFE. This positive effect was further improved with an increasing number of training samples; for example, in the case of 50 training samples per class, the OA improved from 96.6% to 98.4%.

In most cases, the use of the NWFE gave higher accuracies than were obtained for any other FR approach. The use of the NWFE improved the OA accuracy by 1 point to 6 points as compared to the PCA in the case of the University of Pavia data (figure 5). The use of the KPCA also gave significant improvements in terms of accuracies. Using the KPCA with nine bands was better in terms of accuracies than using the PCA with four bands. It is noteworthy that the classification accuracies achieved on different feature sets (nb1, . . ., nb4), generated by the NWFE, were relatively stable; that is, using 25 bands for the University data and 20 bands for the Centre data gave similar accuracies as classifications that were based on seven and five features, respectively. However, using 16 bands of the Centre data with the NWFE gave the highest overall classification accuracy.

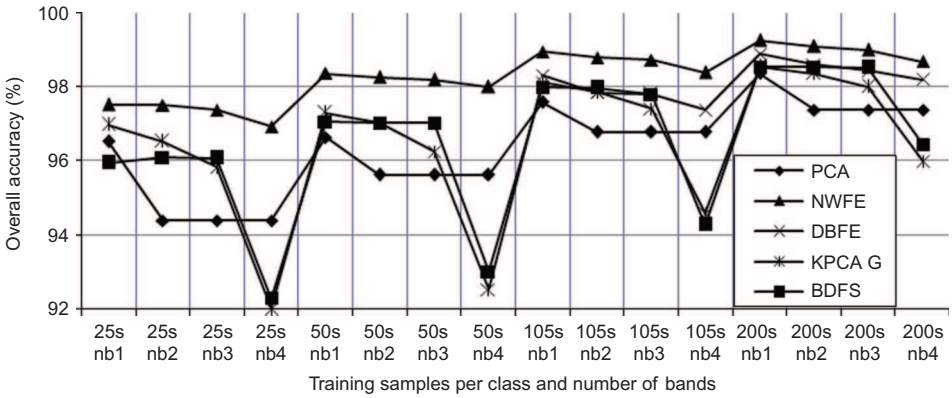


Figure 4. Overall accuracy for the RF classifier and the Centre of Pavia data set, using different numbers of training samples per class and different numbers of bands.

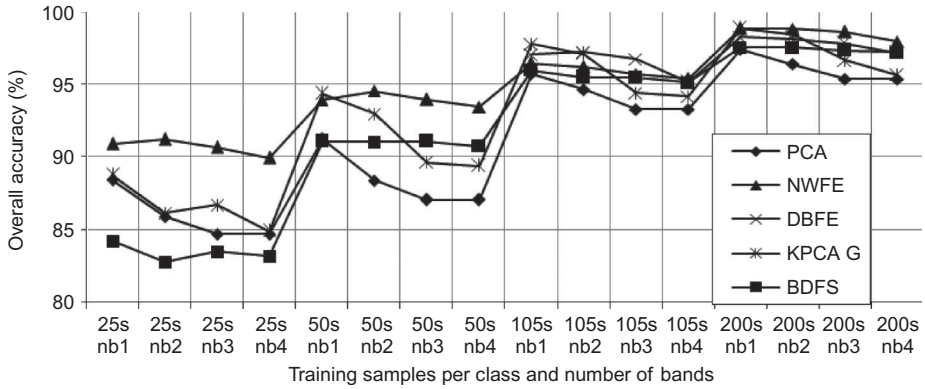


Figure 5. Overall accuracy (OA) for RF classifier and the University of Pavia data set, using different numbers of training samples per class and different numbers of bands.

It could be argued that the above does not give a fair comparison because only a few principal components were used in the experiments and many more channels were used in the classification of the other FR methods. Because of this, additional experiments were run to emphasize that a larger number of bands from the PCA would not further improve the classification accuracies. In these additional experiments, using a higher number of principal components from the PCA gave lower accuracies than using the number of bands corresponding to 99% of the cumulative variance. For instance, when the number of bands increased, the global accuracy fell from 91.2% to 89.7% in the case of the University of Pavia data when 25 training samples per class were used with the SVM.

Table 3 shows the OA, AA and class accuracies for 25 training samples per class for the University data. Comparing the classification of the PCA with that of the NWFE (using 16 bands), the AA remained almost the same but the OA improved from 88.4% to 91.2%. Whereas some classes were classified more accurately using the NWFE (e.g. bare soil, meadows), other classes were classified more accurately using the PCA (e.g. asphalt, bitumen and bricks).

5.3.2 Classifications with the SVM. The results of the classification with the SVM are shown in figures 6 and 7. The experimental results show that the BDFS seems more suitable in terms of accuracy and achieves the highest OA. Comparing the BDFS results with 25 training samples per class to the classification results achieved with the PCA, the OA increased by 7 points (from 85.4% to 92.3%). Furthermore, it is noteworthy that the SVM classification performed well with the KPCA and achieved an OA of 91%.

In the case of the SVM classification of the Centre of Pavia data, the NWFE performed slightly better than the BDFS. The classification of both the NWFE and the BDFS improved the accuracy of the classification for the PCA. The enhancement is between 0.5 and 1 point in every case. The KPCA showed no enhancement when compared to the PCA in this case.

With regard to the class accuracies achieved by the SVM, the BDFS outperforms the other FR techniques in the majority of cases (see table 4). Only in the case of the asphalt class did the use of the PCA result in a higher class accuracy. Comparing the

Table 3. OA, AA and class accuracies for RF and University of Pavia data, using 25 training samples per class and different FE methods (PCA and NWFE).

	PCA 4 bands	NWFE 25 bands	NWFE 16 bands
OA (%)	88.4	90.9	91.2
AA (%)	93.6	93.8	93.9
Class accuracy (%)			
Asphalt	95.6	92.1	91.7
Meadows	81.3	87.2	87.7
Gravel	89.3	84.6	84.1
Trees	95.2	96.4	95.4
Metal sheets	99.6	100	99.9
Bare soil	88.5	96.3	97.2
Bitumen	98.9	96.0	96.0
Bricks	94.6	92.0	93.5
Shadows	99.8	99.9	99.9

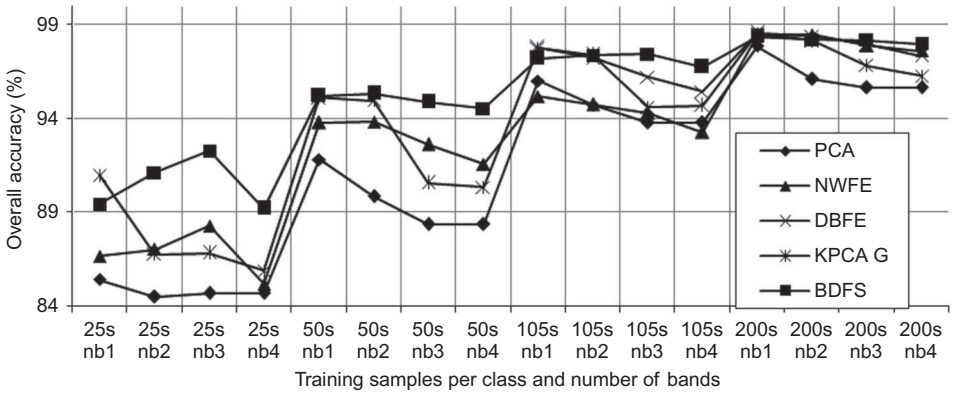


Figure 6. Overall accuracy for the SVM classifier and the University of Pavia data set, using different numbers of training samples per class and different numbers of bands.

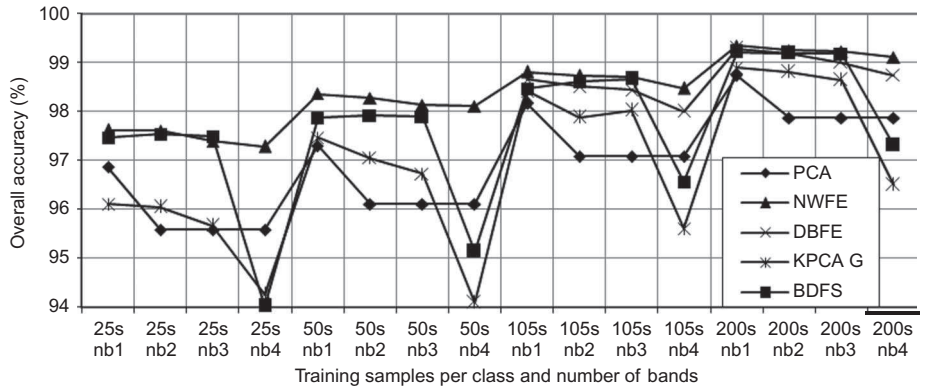


Figure 7. Overall accuracy for the SVM classifier and the Centre of Pavia data set, using different numbers of training samples per class and different numbers of bands.

Table 4. OA, AA and class accuracies for SVM and University of Pavia data, using 25 training samples per class and different FE methods (PCA, NWFE and BDFS).

	PCA 4 bands	NWFE 10 bands	BDFS 15 bands
OA (%)	85.4	88.3	92.3
AA (%)	92.6	92.1	94.8
Class accuracy (%)			
Asphalt	95.3	88.4	92.1
Meadows	75.7	84.8	90.7
Gravel	92.1	86.6	93.5
Trees	96.2	97.0	96.8
Metal sheets	99.6	99.6	99.6
Bare soil	85.5	88.3	89.1
Bitumen	97.8	94.6	97.9
Bricks	90.8	90.0	93.9
Shadows	100.0	100.0	100.0

results achieved with the BDFS, the AA is increased from 92.6% and 92.1% to 94.8%, compared to the results achieved with PCA and NWFE, respectively.

The SEs that were used to build the EMP in the experiments reported above were disc shaped. It is possible that another SE shape might have improved the extraction of the spatial information. To investigate this, the experiments were repeated with square-shaped SEs. We found that the accuracies of the classifications using the square-shaped SEs were lower than in any case when disc-shaped SEs were used to build the EMP, as expected. A possible explanation for this result is that the non-isotropy of the square-shaped SEs may be inadequate.

6. Discussion

The experimental results show that the RF and SVM perform differently with the FR approaches considered. According to the results obtained with the RF, it may be near optimal to select from 10 to 15 features (corresponding to a cumulative variance between 80% and 95% in most cases) in the case of the NWFE. With 25 training samples per class and an RF classifier, selecting 11 features from the Centre data or 16 features from the University data gave the best accuracies.

According to the results obtained with the SVM, selecting from 15 to 20 features (corresponding to an average interclass BD from 1 to 2) showed the best classification enhancement in most of the observed cases. To handle the band selection, a cross-validation step may be performed to determine the optimum number of bands to use.

With 25 training samples per class, the RF classification of the NWFE set improved the OA from 1 to 1.44 points on the Centre of Pavia data and from 2.5 to 7 points on the University of Pavia data compared to classification of the PCA, KPCA and BDFS sets. By contrast, the SVM classification of the BDFS set improved the OA from 0.6 to 1.83 points on the Centre of Pavia data compared to classification of the PCA and KPCA sets (no improvement comparing to the classification of the NWFE set), and from 1.3 to 6.8 points on the University of Pavia data compared to the classification of the PCA, KPCA and NWFE sets.

The classification of the University of Pavia data is much improved by the use of either NWFE or BDFS (depending on the classifier) compared with the classification

of the Centre of Pavia data, meaning the urban density might have a strong effect on the separability between information classes. Nevertheless, the RF classifier performs more accurately using the NWFE method, whereas the BDFS seems more suitable for the SVM classification.

Although the classifications based on the PCA were less accurate, the PCA has one main advantage: most variance is included in only a few bands (i.e. four) and most of it is contained in the first component. By contrast, the other methods require more bands to generate adequate classification results, with the consequent use of more memory. For instance, the use of the NWFE increases the OA of the classification of the University of Pavia with the SVM by 2.8 points, but also generates 16 features, that is to say using the NWFE requires about four times more memory than using the PCA in this case.

The KPCA gave an improvement of the PCA in terms of classification accuracies in every case except for the classification of the Centre of Pavia with the SVM and 25 training samples per class. However, the KPCA uses a few more bands than the PCA because the output of a KPCA algorithm should have the same number of bands as the classes of interest (Fauvel *et al.* 2009). Moreover, it is computationally complex and often needs a validation step to determine the best parameter(s) for the core function. As a consequence, the KPCA computation time is about 10 to 100 times longer than the other FR methods. However, the accuracy enhancement due to the use of the KPCA method is good using either the RF or the SVM.

The DBFE showed good accuracies in many cases but it needs many training samples for processing. This drawback reduces the applicability of the DBFE to the case where a large number of training samples is available.

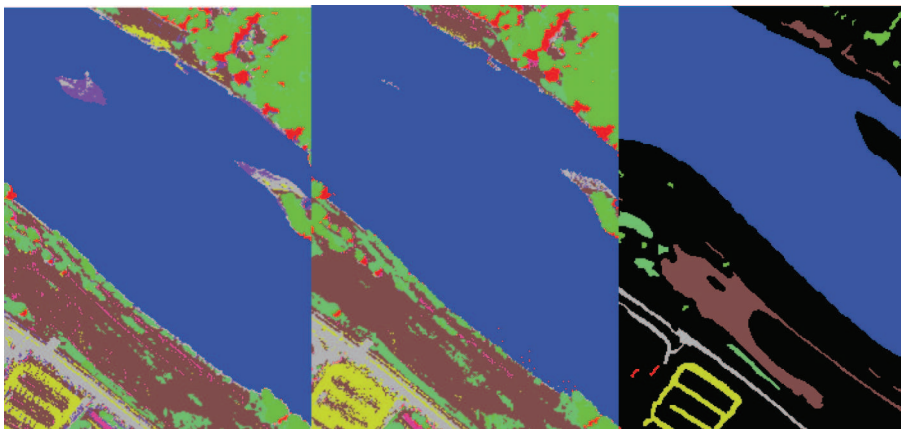
The NWFE is the FR method that shows the best improvement in classification accuracy when the RF classifier is used. Although this method is in many ways similar to the DBFE method and needs several training samples to be applied, it can be processed with very few training samples per class. The NWFE appears to have improved the OA from 0.5 to 1 point in most of the cases (figures 4 and 5) as compared to any other FR method when the RF classifier was used. Furthermore, it seems to have enhanced the classification of the SVM. Nevertheless, the DBFS seems to be slightly more suitable for improvements in the SVM accuracies. However, these differences need to be considered in additional experiments.

Figures 8–10 present a visual comparison of the classification maps, showing the impact of the different FE methods on the final result. In figure 8, the classification map obtained with the PCA appears noisier than the NWFE result and it can be seen that the PCA generates some errors within the water class. In that case, using the NWFE instead of the PCA improved the OA from 96.2% to 98.3%. Furthermore, the class accuracy for water increased from 98.7% to 99.9% and for asphalt it increased from 88.6% to 96.2%.

The meadow and bare-soil classes were not very well discriminated when the PCA was used (figure 9). The use of the NWFE tackled this problem. Thus, the OA accuracy increased 91.2% from to 93.2%, the meadow class accuracy from 86.1% to 91.4% and the bare-soil class accuracy from 89.3% to 94%.

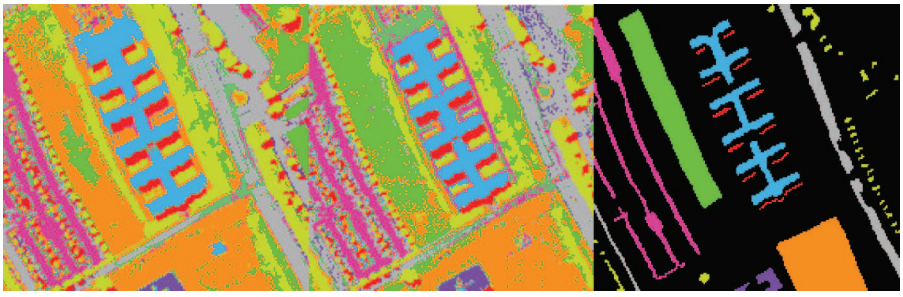
The DBFS shows the best accuracy improvements in the case of a classification with SVM, especially with only a few training samples per class. On average, using the DBFS instead of any other FE method improved the accuracy by up to 3 points in the case of 25 samples per class.

Figure 10 shows a missed square of meadows using the PCA preprocessing, whereas it is correctly classified using the BDFS. The BDFS appears to perform well especially



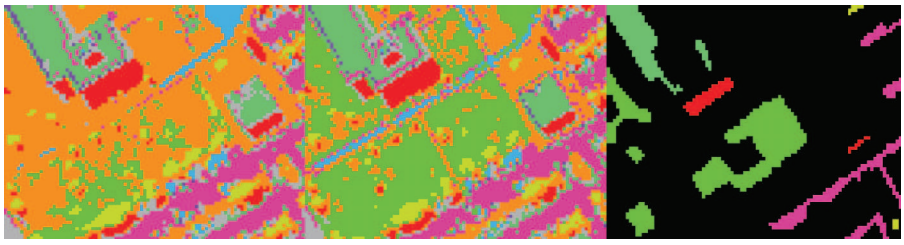
(*a*) (*b*) (*c*)

Figure 8. Partial classification maps of the Centre of Pavia obtained from the RF classifier and 50 training samples per class: (*a*) PCA; (*b*) NWFE; (*c*) ground truth.



(*a*) (*b*) (*c*)

Figure 9. Partial classification maps of the University of Pavia obtained from the SVM classifier and 50 training samples per class: (*a*) PCA preprocessing; (*b*) NWFE preprocessing; (*c*) ground truth map.



(*a*) (*b*) (*c*)

Figure 10. Partial classification maps of the University of Pavia obtained from an SVM classifier and 25 training samples per class: (*a*) PCA preprocessing; (*b*) BDFS preprocessing; (*c*) ground truth map.

on pixels not belonging to any building. In this case, using the BDFS instead of the PCA improved the OA from 87.4% to 92.9%, the meadow class accuracy from 79.9% to 91.6% and the bare-soil class from 85.1% to 91.1%.

7. Conclusion

In this study we investigated the influence of different FR methods on the classification accuracy of an EMP for hyperspectral remote sensing data. Two classifiers, the RF and the SVM, were used in the classification. The classifiers showed different characteristics in terms of classification accuracies for different FR methods. However, the results show that, in many cases, PCA is inadequate for the construction of an EMP in terms of classification accuracy. In the classification of an EMP with the RF approach, the best results were obtained in combination with the NWFE. By contrast, the BDFS performed well with the SVM in terms of accuracy. Nevertheless, the PCA still has some advantages; in particular, the computational time is short in comparison to the BDFS. Moreover, an appropriate classification based on PCA requires a much smaller number of features than the NWFE and the BDFS. Thus, the PCA again requires a lower computation time. However, with an increased computational power currently being available and the introduction of parallel implementations, the increased computational time of the NWFE and BDFE is not a crucial issue. Overall, the experimental results indicate that the choice of the FR method used to generate an EMP affects the final classification accuracy. Furthermore, methods other than the PCA tend to be more adequate for preprocessing of the EMP in terms of classification accuracy. However, the differences between the methods need to be considered further in additional experiments in future work.

Acknowledgements
This research was supported in part by the Research Fund of the University of Iceland and the Icelandic Research Fund.

References

Akcay, H.G. and Aksoy, S., 2008, Automatic detection of geospatial objects using multiple hierarchical segmentations. *IEEE Transactions on Geoscience and Remote Sensing*, **46**, pp. 2097–2111.

Bellens, R., Gautama, S., Martinez-Fonte, L., Philips, W., Chan, J.C.-W. and Canters, F., 2008, Improved classification of VHR images of urban areas using directional morphological profiles. *IEEE Transactions on Geoscience and Remote Sensing*, **46**, pp. 2803–2813.

Benediktsson, J.A., Palmason, J.A. and Sveinsson, J.R., 2005, Classification of hyperspectral data from urban areas based on extended morphological profiles. *IEEE Transactions on Geosciences and Remote Sensing*, **43**, pp. 309–320.

Benediktsson, J.A., Pesaresi, M. and Arnason, K., 2003, Classification and feature extraction for remote sensing images from urban areas based on morphological transformations. *IEEE Transactions on Geosciences and Remote Sensing*, **41**, pp. 1940–1949.

Bernard, S., Heutte, L. and Adam, S., 2009, Influence of hyperparameters on random forest accuracy. *Lecture Notes in Computer Science*, **5519**, pp. 171–180.

Breiman, L., 2001, Random forests. *Machine Learning*, **45**, pp. 5–32.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J., 1984, *Classification and Regression Trees* (Belmont: Wadsworth International Group).

Bruzzone, L. and Persello, C., 2009, A novel context-sensitive semisupervised SVM classifier robust to mislabeled training samples. *IEEE Transactions on Geoscience and Remote Sensing*, **47**, pp. 2142–2154.

Burges, C., 1998, A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, **2**, pp. 121–167.

Camps-Valls, G., Gomez-Chova, L., Munoz-Mari, J., Rojo-Alvarez, J.L. and Martinez-Ramo, M., 2008, Kernel-based framework for multitemporal and multisource remote sensing data classification and change detection. *IEEE Transactions on Geosciences and Remote Sensing*, **46**, pp. 1822–1835.

Camps-Valls, G., Gomez-Chova, L., Munoz-Mari, J., Vila-Frances, J. and Calme-Maravilla, J., 2006, Composite kernels for hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters*, **3**, pp. 93–97.

Chi, M. and Bruzzone, L., 2007, Semisupervised classification of hyperspectral images by SVMs optimized in the primal. *IEEE Transactions on Geosciences and Remote Sensing*, **45**, pp. 1870–1880.

Cristianini, N. and Shawe-Taylor, J., 2000, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods* (Cambridge: Cambridge University Press).

Fauvel, M., Benediktsson, J.A., Chanussot, J. and Sveinsson, J., 2008, Hyperspectral data using SVMs and morphological profiles. *IEEE Transactions on Geosciences and Remote Sensing*, **46**, pp. 3804–3814.

Fauvel, M., Chanussot, J. and Benediktsson, J.A., 2009, Kernel principal component analysis for the classification of hyperspectral remote sensing data over urban areas. *EURASIP Journal on Advances in Signal Processing*, **2009**, 783194, doi:10.1155/2009/783194.

Ghoggali, N., Elgani, F. and Bazi, Y., 2009, A multiobjective genetic SVM approach for classification problems with limited training samples. *IEEE Transactions on Geosciences and Remote Sensing*, **47**, pp. 1707–1718.

Gislason, P.O., Benediktsson, J.A. and Sveinsson, J.R., 2006, Random forests for land cover classification. *Pattern Recognition Letters*, **27**, pp. 294–300.

Ham, J., Chen, Y. and Crawford, M.M., 2005, Investigation of the random forest framework for classification of hyperspectral data. *IEEE Transactions on Geosciences and Remote Sensing*, **43**, pp. 492–501.

Herold, M., Roberts, D.A., Gardner, M.E. and Philip, E.D., 2004, Spectrometry for urban area remote sensing: development and analysis of a spectral library from 350 to 2400 nm. *Remote Sensing of Environment*, **91**, pp. 304–319.

Huang, C., Davis, L.S. and Townshend, J.R., 2002, An assessment of support vector machines for land cover classification. *International Journal of Remote Sensing*, **23**, pp. 725–749.

Joelsson, S.R., Benediktsson, J.A. and Sveinsson, J.R., 2005, Random forest classifiers for hyperspectral data. *IEEE International Geoscience and Remote Sensing Symposium. IGARSS '05. Proceedings. 2005 IEEE International*, pp. 160–163.

Kuo, B.C. and Landgrebe, D.A., 2004, Nonparametric weighted feature extraction for classification. *IEEE Transactions on Geosciences and Remote Sensing*, **42**, pp. 1096–1105.

Landgrebe, D.A., 2002, *Signal Theory Methods in Multispectral Remote Sensing* (New York: John Wiley & Sons).

Lee, C. and Landgrebe, D.A., 1993, Feature extraction based on decision boundaries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **15**, pp. 388–400.

Lee, C. and Landgrebe, D.A., 1997, Decision boundary feature extraction for neural networks. *IEEE Transactions on Neural Networks*, **8**, pp. 75–83.

Matheron, G., 1967, *Eléments pour une Théorie des Milieux Poreux* (Paris: Masson).

Matheron, G., 1975, *Random Sets and Integral Geometry* (New York: Wiley).

Pal, M. and Mather, P.M., 2006, Some issues in the classification of DAIS hyperspectral data. *International Journal of Remote Sensing*, **27**, pp. 2895–2916.

Pesaresi, M. and Benediktsson, J.A., 2001, A new approach for the morphological segmentation of high-resolution satellite imagery. *IEEE Transactions on Geosciences and Remote Sensing*, **39**, pp. 309–320.

Polikar, R., 2006, Ensemble-based systems in decision making. *IEEE Circuits and Systems Magazine*, **6**, pp. 21–45.

Richards, J.A. and Jia, X., 1999, *Remote Sensing Digital Image Analysis: An Introduction* (Berlin: Springer).

Serra, J., 1982, *Image Analysis and Mathematical Morphology*, vol. I (London: Academic Press).

Serra, J., 1988, *Image Analysis and Mathematical Morphology: Theoretical Advances*, vol. II (London: Academic Press).

Soille, P., 2003, *Morphological Image Analysis: Principles and Applications*, 2nd edn (Berlin: Springer-Verlag).

Soille, P. and Pesaresi, M., 2002, Advances in mathematical morphology applied to geosciences and remote sensing. *IEEE Transactions on Geosciences and Remote Sensing*, **40**, pp. 2042–2055.

Tuia, D., Pacifici, F., Kanevski, M. and Emery, W.J., 2009, Classification of very high spatial resolution imagery using mathematical morphology and support vector machines. *IEEE Transactions on Geoscience and Remote Sensing*, **47**, pp. 3866–3879.

Vapnik, V.N., 1998, *Statistical Learning Theory* (New York: John Wiley & Sons).

Waske, B., Benediktsson, J.A., Arnason, K. and Sveinsson, J.R., 2009, Mapping of hyperspectral AVIRIS data using machine learning algorithms. *Canadian Journal of Remote Sensing*, **35**, pp. 106–116.