Automatic Spectral-Spatial Classification Framework based on Attribute Profiles and Supervised Feature Extraction

Pedram Ghamisi, *Student Member, IEEE,* Jon Atli Benediktsson, *Fellow, IEEE* Johannes R. Sveinsson, *Senior Member, IEEE,*

***Abstract*—A robust framework for classification of hyperspectral images which takes into account both spectral and spatial information is proposed. Extended Multivariate Attribute Profile (EMAP) is used for extracting spatial information. Moreover, for solving the so-called curse of dimensionality, supervised feature extraction is carried out on both the original hyperspectral data and the output of the EMAP. After performing the dimensionality reduction, two output vectors of the original data and attributes are concatenated into one stacked vector. The final classification map is achieved by using a random forest classifier. The main difficulties of using an EMAP is to initialize the attribute parameters. Therefore, a fully automatic scheme of the proposed method is introduced to overcome the shortcomings of using EMAP. The proposed method is tested on two widely-known data sets. Experimental results confirm that the proposed method provides an accurate classification map in an acceptable CPU processing time.**

***Index Terms*—Hyperspectral image analysis, Spectral-spatial Classification, Attribute Profile, Feature extraction, Random Forest Classifier, Automatic Classification.**

I. Introduction

SUPERVISED classification plays a key role in remote sensing image processing and is important in many applications, including crop monitoring, forest applications, urban development, mapping and tracking and risk management. Thanks to recent advances in remote sensing technologies, the spatial resolution of the sensed images has increased. This has led to a better identification of relatively small structures such as roads and houses. Hyperspectral sensors capture hundreds of spectral bands from ultraviolet to infrared for each image pixel, which is helpful for detailed physical analysis of structures in the captured image [1].

In the spectral domain, each spectral channel is considered as one dimension. By increasing the features in the spectral domain, theoretical and practical problems may arise. For instance, while keeping the number of training samples constant, the classification accuracy actually decreases when the number of features becomes large [2]. For the purpose of classification, these problems are related to the curse of dimensionality. In [3], Landgrebe shows that too many spectral bands can be undesirable from the standpoint of expected classification accuracy because the accuracy of the statistics estimation decreases (Hughes phenomenon). The aforementioned issue demonstrates that there is an optimal number of bands for classification accuracy and more features do not necessarily lead to better results. Therefore, use of feature reduction techniques may lead to a better classification accuracy [4].

Conventional spectral classifiers consider the hyperspectral image as a list of spectral measurements with no spatial organization [5]. A joint spectral and spatial classifier is required in order to reduce the labeling uncertainty that exits when only spectral information is taken into account, and helps to overcome the salt and pepper appearance of the classification map. Furthermore, other relevant contextual information can be extracted when the spatial domain is considered. As an example, for a given pixel, it is possible to extract the size and the shape of the structure to which it belongs. Therefore, using a combination of spectral and spatial information can improve the accuracy of the classification.

In order to extract of spatial information, two neighborhood systems are available: Crisp neighborhood system and adaptive neighborhood system. One way for extracting spatial information by using crisp neighbors is to consider Markov Random Field (MRF) modeling. MRF is a family of probabilistic models that can be described as a 2-D stochastic processes over discrete pixels latices[6]. They can be considered as a powerful tool for incorporating spatial and contextual information into the classification framework. There is an extensive literature on the use of MRFs in classification such as [7], [8], [9], [10], [11], [9], [12]. Texture analysis is another way of considering a crisp neighbor system to extract spatial information. That approach is widely used in remote sensing (e.g., in [13]). However, the main shortcoming of considering a set of crisp neighbors is that 1) the standard neighborhood system may not contain enough samples, which decreases the effectivity of the classifier in particular when the input data set is of high resolution and the neighboring pixels are highly correlated [1], and 2) a larger neighborhood system leads to intractable computational problems [1].

To solve the above problem, an adaptive neighborhood system can be taken into account. One way for considering the adaptive neighborhood system is to use different types of segmentation methods. Image segmentation is a procedure which can be used to modify the accuracy of classification maps [14]. To make such an approach effective, an accurate segmentation of the image is needed [15]. An extensive

P. Ghamisi, J. A. Benediktsson and J. R. Sveinsson are with the Faculty of Electrical and Computer Engineering, University of Iceland, 107 Reykjavik, Iceland (corresponding author, e-mail: peg6@hi.is)

This research was supported in part by the Icelandic Research Fund for Graduate Students.

literature is available on the extraction of spatial information using segmentation techniques (e.g., [16], [17] and [18]).

Another set of methods which can extract spatial information by using adaptive neighbor system is based on morphological filters. Pesaresi and Benediktsson [19] used morphological transformations to build a Morphological Profile (MP). In [20], the concept of MP was extended in order to handle hyperspectral images and the extension was named the Extended MP (EMP). In [21], EMP was used along with the Support Vector Machine (SVM) classifier in order to perform spectral and spatial classification on hyperspectral images. Attribute Profile (AP) is another extension of MP, and provides a multilevel characterization of an image by using the sequential application of morphological attribute filters which can be considered for modeling different specifications of the structural information [22]. AP is a powerful tool to increase the discrimination of different classes [22], [23].

In this paper, a new automatic approach is proposed for the accurate classification of remote sensing images. Although the method can be used for the classification of multispectral images with a coarse spectral resolution, it is used here for spectral and spatial classification of hyperspectral images. The spatial part of the method consists of both unsupervised and supervised feature extraction and the EMAP. Supervised feature extraction is applied on the spectral part. Furthermore, the result of the spatial and spectral features are gathered into a stacked vector. In the field of hyperspectral image analysis, the most widely used classifiers are Random Forest (RF) and Support Vector Machine classifiers (SVMs). These two methods are comparable in the sense of classification accuracies. However, while both methods are shown to be effective classifiers for non-linear classification problems, SVM requires an exhaustive (computationally intensive) parameter tuning (e.g., model selection performed on a grid) for optimal results, whereas RF does not require any such tuning. In addition, overall accuracy is not the only critical issue for the purpose of hyperspectral image classification. Another critical index which can evaluate the efficiency of a classifier is the CPU processing time. In this sense, RF is faster than SVM and therefore, we prefer using RF instead of SVM in the final step, where the produced stacked vector is classified. The proposed method is performed in experiments on two well-known data sets. Results confirm that the new method is able to effectively classify hyperspectral images both in terms of classification accuracies and CPU processing time. To make the new approach as efficient as possible, an automatic scheme for the new method is introduced in this paper in order to solve the main difficulties in using the EMAP.

This paper is organized as follows: the proposed methodology is discussed in Section II. Then, Section III is devoted to experimental results. Finally, Section IV outlines the main conclusions.

II. METHODOLOGY

In the proposed method, supervised Feature Extraction (FE) is performed initially on the input data. In parallel, the input data are transformed by Principal Component Analysis (PCA) and the most important Principal Components (PCs) are used as base images for the EMAP. Then, the supervised FE is performed on the output of the EMAP. Furthermore, first features with a cumulative variance more than 99% are selected as the output of the supervised FE. Finally, the spatial and spectral features are gathered into a stacked vector and classified by the RF classifier and a final classification map is achieved. Fig. 1 illustrates the flowchart of the new method. In the following, specific parts of the proposed framework will be discussed in detail.

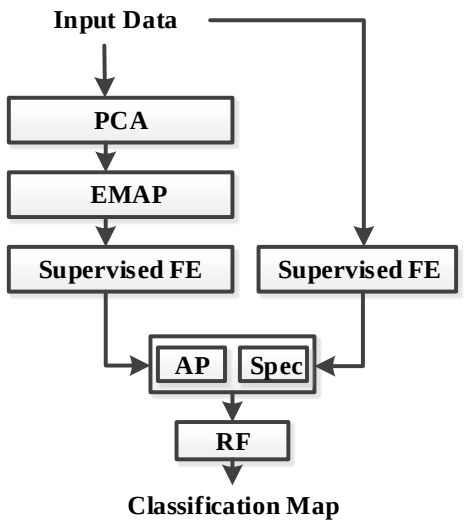

Fig. 1. A flowchart of the proposed method.

A. Feature Extraction (FE)

FE can be explained as finding a set of vectors that represents an observation while reducing the dimensionality. From one point of view, FE can be classified into two categories; unsupervised and supervised FE where the former is used for the purpose of data presentation and latter is considered for solving the so-called Hughes phenomena [2] and reducing the redundancy of data in order to improve classification accuracies. In pattern recognition, it is desirable to extract features which are focused on the discrimination between classes of interest. Although a reduction in dimensionality is of importance, the error rising from the reduction in dimension has to be without sacrificing the discriminative power of classifiers [21]. In this work, PCA is used for the purpose of unsupervised FE and Discriminant Analysis Feature Extraction (DAFE), Decision Boundary Feature Extraction (DBFE) and Nonparametric Weighted Feature Extraction (NWFE) are taken into consideration for the purpose of supervised FE. Below, PCA, DAFE, DBFE and NWFE are described in more detail.

1) PCA: The general aim of PCA is to transform the data into a lower dimensional subspace which is optimal in terms of sum-of-squared error [24]. PCA reduces the dimensionality of

a data set with interrelated variables, while retaining as much as possible of the variation in the data set. The dimensionality reduction is obtained by a linear transformation of the data into a new set of variables, the PCs. The PCs are orthogonal to each other and are ordered in such a fashion that the first PC corresponds to the greatest variance, the second component corresponds to the second greatest variance and so on.

2) DAFE: This approach is widely used for dimension reduction in classification problems [25]. Since, DAFE uses the mean vector and the covariance matrix of each class, it is considered as supervised feature extraction. In DAFE, within-class, between-class and mixture scatter matrices are usually considered as the criteria of class separability. DAFE is fast and works well when the distribution of data is normal. Otherwise, the performance of DAFE may not be satisfactory. Another problem associated with this method is that if the difference in the class-mean vectors is small, the feature chosen will not be reliable. In the same way, if one class-mean vector is very different from others, its class will eclipse the others in the computation of the between-class covariance matrix [26]. As a consequence, the feature extraction process will be ineffective. In addition, DAFE performs the computations at full dimensionality, which requires a large number of training samples in order to accurately estimate parameters. The main shortcoming of DAFE is that DAFE is not full rank but its rank at maximum is equal to L-1 where L is the number of classes. Assume the rank of the within-class scatter matrix is u, then only $min(L-1, u)$ features are selected by using DAFE. Since in real situations, the data distribution is complicated, using only L-1 features usually is not sufficient.

3) DBFE: This method was proposed in [27] where it was shown that both discriminantly informative features and redundant features can be extracted from the decision boundary between two classes. The features are extracted from the Decision Boundary Feature Matrix (DBFM). In order to obtain the same classification accuracy as in the original space, keeping the eigenvectors of the decision boundary feature matrix corresponding to nonzero eigenvalues is crucial. The performance of this method does not deteriorate even when there is no difference in the mean vectors or the covariance matrices and the approach does not rely on the number of classes in the same way as the DAFE. The efficiency of DBFE is highly dependent on training samples which is not desirable. Another shortcoming of DBFE is that it can be computationally intensive.

4) NWFE: In order to overcome the limitations of DAFE and DBFE, NWFE was introduced in [28]. NWFE is developed based on DAFE by focusing on samples near the eventual decision boundary. The main ideas behind NWFE are to put different weights on different samples in order to compute 'weighted means' and define new nonparametric within-class and between-class scatter matrices.

B. Extended Multivariate Attribute Profile (EMAP)

1) Connected components: A connected component is regarded as a group of *iso*-level pixels which are connected according to a predefined connectivity rule. Two pixels are connected based on a connectivity rule. The most well-known connectivity rules are 4- and 8-connected, where a pixel is considered as adjacent to the four or eight of its neighboring pixels, respectively.

2) Basic morphology operators: Erosion and dilation are considered as the alphabets of mathematical morphology. These operators are performed on an image with a set of known shape, called a Structuring Element (SE). Opening and closing are combinations of erosion and dilation. These operators simplify the input image by removing structures with size less than the SE. However, these operators make changes on the shape of the structures which are still present in the image after the opening/closing. Therefore, they can introduce fake objects in the image [21]. One way to solve this issue is to consider opening and closing by reconstruction. Opening and closing by reconstructions are connected operators that satisfy the following criterion: If the SE cannot fit the structure of the image, then it is totally removed, otherwise it is totally preserved. Reconstruction operators remove objects smaller than the SE without altering the shape of those objects and reconstruct connected components from the preserved objects. For gray scale images, opening by reconstruction removes unconnected light objects and in dual, closing by reconstruction removes unconnected dark objects.

C. Morphological Profile (MP)

To determine the shape or size of all structures present in an image, it is crucial to use a range of different SE sizes for the better analysis of structures in the image. MPs are defined using successive opening/closing operations with an SE of an increasing size. The successive usage of opening/closing leads to a simplification of the input image and a better understanding of different structures in the image. An MP is composed of the opening profile and the closing profile. Although MP is a powerful tool for the extraction of spatial information, it suffers by few limitations such as:

1) The shape of SEs is fixed which is considered as a main limitation for the extraction of objects within a scene.
2) SEs are unable to describe information related to the gray-level characteristics of regions such as spectral homogeneity, contrast and so on.
3) A final limitation associated with the concept of MPs is the computational complexity. The original image needs to be completely processed for each level of the profile, which demands two complete processings of the image; one performed by a closing transformation and the other by an opening transformation. Thus, the complexity increases linearly with the number of levels included in the profile [22].

D. Attribute Profile (AP)

A morphological attribute profile is considered as the generalization of the MP which provides a multilevel characterization of an image by using the sequential application of morphological attribute filters [22]. Morphological attribute opening and thinning are morphological Attribute Filters (AFs) which were introduced in [29]. AFs are connected operators

which process an image by considering only its connected components. For binary images, the connected components are simply the foreground and background regions present in the image. In order to deal with gray scale images, the set of connected components can be obtained by considering the image to be composed by a stack of binary images generated by thresholding the image at all its grey-level values [30].

AFs process an image based on a given criterion. AFs keep or merge the connected component CC_i based on a logical predicate, T, if a given attribute is greater/lower than an arbitrary reference, such as $T^a_\kappa(CC_i) = a(CC_i) > \kappa$ where a is an attribute and κ is an arbitrary reference value [23]. The criterion is evaluated on all the connected components of the image and if the criterion is not met, the region is merged to the adjacent region with a closer gray-level value. If the regions with lower (greater) gray level values are taken into account in the merging process, then the transformation is considered as anti-extensive (extensive) [22]. The transformation is idempotent if the result of transformation is not dependent on the number of times a transformation with the same parameter is performed. A transformation with the aforementioned specifications is called thinning (thickening).

An AP is obtained by the sequence of attribute thinning and thickening transformations defined with a sequence of progressively stricter criteria [22]. Let ϕ^κ and γ^κ be an attribute thickening and attribute thinning, respectively. AP of the image f with the set of N criteria is shown by $AP(f)$. Mathematically, $AP(f)$ can be expressed as

$$AP(f) = \{\phi^{\kappa_N}(f), \phi^{\kappa_{N-1}}(f), ..., \phi^{\kappa_1}(f), f, \gamma^{\kappa_1}(f), ..., \gamma^{\kappa_{N-1}}(f), \gamma^{\kappa_N}(f)\}. \quad (1)$$

To handle hyperspectral images, the extension of AP was proposed in [31]. Extended AP (EAP) is a stacked vector of different APs computed on the first C features extracted from the original data set (I^D with D dimensions) and fe shows a feature. EAP is given by

$$EAP(I^D) = \left\{AP(fe_1(I^D)), AP(fe_2(I^D)), ..., AP(fe_C(I^D))\right\}. \quad (2)$$

When concatenation of different attributes, $a_1, a_2, ..., a_M$ are gathered into a stacked vector, the EMAP is obtained [31] and is given mathematically by

$$EMAP(I^D) = \left\{EA\acute{P}_{a_1}(I^D), EA\acute{P}_{a_2}(I^D), ..., EA\acute{P}_{a_M}(I^D)\right\}, \quad (3)$$

where $EA\acute{P} = EAP\{PC_1, ..., PC_c\}$ and a_i is a generic attribute.

The application of the profiles for large volumes of data is computationally demanding and that is considered to be one of the main difficulties in using them. In order to solve this issue, the efficient implementation of attribute filters was proposed in [32]. Salembier *et al.* in [32], introduced a new data representation named Max-tree which has received much interest since it increases the efficiency of filtering by dividing the transformation process into three steps: 1) tree creation; 2) filtering; and 3) image restitution [22].

E. Fusion of extracted features via vector stacking

As can be seen from Fig. 1, the original data are transformed by a Supervised FE approach in order to provide a few effective features that contain the spectral information of the input data. Let ζ_φ be as the features associated with the spectral bands.

With reference to Fig. 1, the input data are transformed by PCA and the first effective PCs are used in order to reduce the redundancy in the data but keeping most of the variation. Then, EMAP is computed by using only the first effective PCs that correspond to the 99% of the variance. Afterward, each AP is composed of n thickening and n thinning transformations of the corresponding PC for each attribute. In order to produce the Multivariate AP (MAP) for each PC, depending on the number of attributes (e.g., m different attributes), we come up with $m(2n)+1$ number of features in each MAP. Finally, the number of features in the EMAP by considering P PCs is equal to $P(m(2n)+1)$. Let ζ_ω be as the features associated to the EMAP. Finally, the obtained stack vector is $\zeta = [\zeta_\varphi, \zeta_\omega]^T$.

F. Random Forest (RF)

RF was first introduced in [33] and is an ensemble method for classication and regression. Ensemble classiers get their name from the fact that several classiers, i.e., an ensemble of classiers, are trained and their individual results are then combined through a voting process. For the purpose of classification of an object from an input vector, the input vector is run down each tree in the forest. Each tree provides a unit vote for a particular class and the forest chooses the classication having the most votes. Based on studies in [34], the computational complexity of the RF algorithm is $cT\sqrt{MN}log(N)$ where c is a constant, T denotes the number of trees in the forest, M is regarded as the number of variables and N is the number of samples in the data set. It is easy to detect that RF is not computationally intensive but demands a considerable amount of memory since it needs to store an N by T matrix while running. RF can provide a good classification result in terms of accuracies and does not assume any underlying probability distribution for input data. Another advantage of RF classifier is that it is insensitive to noise in the training labels. In addition, RF provides an unbiased estimate of the test set error as trees are added to the ensemble and finally it does not overfit.

G. Automatic Scheme for EMAP

The main difficulties of using the EMAP are 1) to know which attributes lead to a better discrimination for different classes, and 2) which threshold values should be considered in order to initialize each AP. In this section, an automatic scheme of the proposed method is introduced in order to solve the latter problem. While the APs can be constructed by using a wide variety of attributes, in the automatic scheme, the area and standard deviation attributes are only used, since the aforementioned attributes can be adjusted in an automatic way and are well-related to the object hierarchy in the images. The standard deviation is adjusted with respect to the mean

of the individual features, since the standard deviation shows dispersion from the mean [35]. Therefore, λ_s is initialized in a fashion to cover a reasonable amount of deviation in the individual feature, which is mathematically given by

$$\lambda_s(PC_i) = \frac{\mu_i}{100}\{\sigma_{min}, \sigma_{min} + \delta_s, \sigma_{min} + 2\delta_s, ..., \sigma_{max}\}, \tag{4}$$

where μ_i is the mean of the i-th feature and σ_{min}, σ_{max} and δ_s are 2.5, 27.5 and 2.5 percent, respectively, which leads to 11 thinning and 11 thickening operations.

With regard to adjusting λ_a for the area attribute, the resolution of the image should be taken into account in order to construct EAP [23]. The automatic scheme of the attribute area is given below:

$$\lambda_a(PC_i) = \frac{1000}{\upsilon}\{a_{min}, a_{min} + \delta_a, a_{min} + 2\delta_a, ..., a_{max}\}, \tag{5}$$

where a_{min} and a_{max} are initialized by 1 and 14, respectively, with a step increase δ_a equal to 1 and υ shows the spatial resolution of the input data. The EAP for the area attribute includes 14 thinning and 14 thickening operations for each feature. Each level is provided in square meters by considering the resolution of the image υ in meters. Each profile covers structures in the range of 1000 to 14000 m^2, which might be a reasonable range of sizes for different structures in both urban and rural cases in remote sensing images [23]. However, different ranges can be considered for different applications. It should be noted that the aforementioned parameters have been tested on other well-known data sets such as Indian Pines in [23] and results show that these parameters are data set distribution independent and can provide excellent results in terms of classification accuracies. In the introduced framework, one only needs to establish a range of parameter values in order to automatically obtain a classification result with high accuracy for different data sets.

Fig. 2 shows the general idea of the automatic scheme of the proposed method. First, the input data are transformed via PCA and the first PCs with a cumulative variance of more than 99% are kept since they provide the most of data variation. Then, MAP including area and standard deviation attributes with respect to (4) and (5) are built for each PC. Furthermore, the MAP of different PCs are concatenated into a stacked vector. Finally, in order to extract spatial information, the stacked vector is transformed by a supervised FE, the first features with cumulative variance more than 99% are kept, and ζ_ω is the output of this step. In parallel, in order to provide the spectral information, a Supervised FE is performed on the input data set, and ζ_φ is the output of this step. The final classification map is provided by performing RF on the output of the stack vector, $\zeta = [\zeta_\varphi, \zeta_\omega]^T$.

III. EXPERIMENTAL RESULTS

A. Data Description

The test cases are hyperspectral data sets which were captured of the city of Pavia, Italy by Airborne data from the ROSIS-03 (Reflective Optics System Imaging Spectrometer). The ROSIS-03 sensor has 115 data channels with a spectral coverage ranging from 0.43 to 0.86 μm. The data have been atmospherically corrected, but not geometrically corrected. The spatial resolution is 1.3 m per pixel.

1) Pavia University: The first data set is of the Engineering School at the University of Pavia and consists of different classes including: trees, asphalt, bitumen, gravel, metal sheet, shadow, bricks, meadow and soil. In the experiments, 12 noisy data channels are eliminated and 103 data channels are used for processing. The original data set is 610 by 340 pixels. Fig. 3. (a) shows a false color composite of the Pavia University scene. The available test and training samples are listed in Table. I.

2) Pavia Center: The second data set is captured of the center of Pavia. The original data set is 1096 by 1096 pixels. A 381-pixel-wide black stripe in the left part of the data set was removed, leading to 1096 by 715 pixels. Thirteen data channels were removed due to the noise and 102 bands were processed. This data set consists of nine classes, i.e., water, trees, meadows, bricks, soil, asphalt bitumen, tiles, and shadows. Fig. 3. (b) depicts a false color composite of the Pavia Center. The available test and training samples are listed in Table. IV.

B. General Information

The input image is transformed by PCA and the first PCs with a cumulative variation of more than 99 percent are kept, since they contain almost all of the variance in the data sets. For Pavia University, three PCs are necessary to retain 99% of the variance criterion (the cumulative sum of eigenvalues in percentage is 99.55%). In the same way, for Pavia Center, three PCs are necessary to retain 99% of the variance criterion (the cumulative sum of eigenvalues in percentage is 99.47%). In the same fashion, for all supervised FE (DAFE, DBFE and NWFE), the first features with a cumulative variance of more than 99 percent are selected. The number of trees for RF classifier is equal to 200.

In this paper, Spectral refers to when only spectral information is classified by RF. In the same way, when only spatial information are taken into account, it is called AP. AP + Spectral is referred to the classification of the stacked vector including both spectral and spatial information without performing feature reduction. For simplification, the proposed approach using DAFE, DBFE and NWFE are called DAFE, DBFE and NWFE, respectively.

For the MANUAL scheme of the proposed method, four attributes are considered

1) (a) area of the region (related the size of the regions)
2) (s) standard deviation (as an index for showing the homogeneity of the regions)
3) (d) diagonal of the box bounding the regions
4) (i) moment of inertia (as an index for measuring the elongation of the regions).

The values of each attribute are adjusted based on studies in [31] and given below:
$\lambda_a = \{100, 500, 1000, 5000\}$
$\lambda_s = \{20, 30, 40, 50\}$
$\lambda_d = \{10, 25, 50, 100\}$

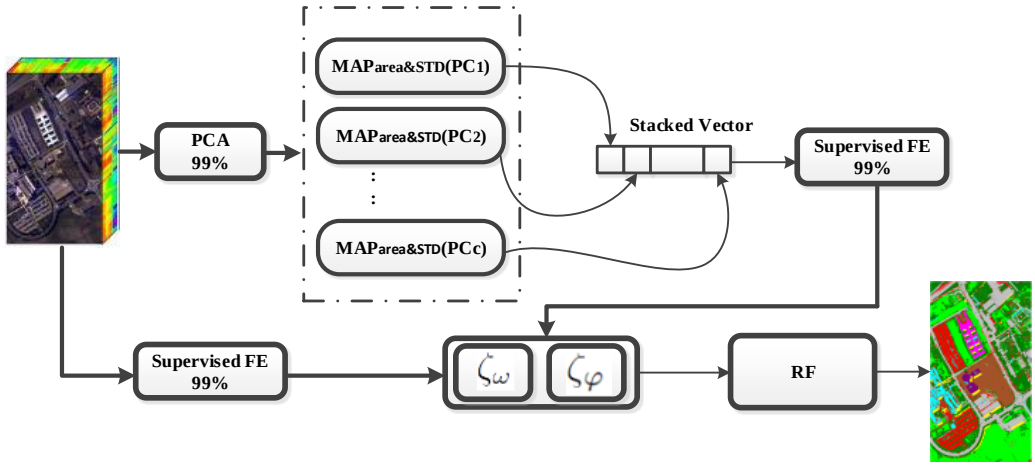

Fig. 2. The general idea of the AUTOMATIC scheme of the proposed method. First, PCA is performed on the input data and first PCs with a cumulative variance of more than 99% are kept. Then, MAP including area and standard deviation attributes with respect to (4) and (5) are built for each PC. Furthermore, the MAP of different PCs are concatenated into a stacked vector. After that, the supervised FE is carried out on the stacked vector and first features with a cumulative variance of more than 99% are kept. The output of this part provides the spatial information of the method. In parallel, the supervised FE is performed on the input data, and first features with a cumulative variance of more than 99% are selected. The output of this step is considered as spectral information As the last stage, the spectral and spatial information are concatenated in a stacked vector and the stacked vector is classified by RF.

Fig. 3. ROSIS data: (a) University and (b) Center. Data specifications are detailed in Tables I and Table IV.

$\lambda_i = \{0.2, 0.3, 0.4, 0.5\}$

As mentioned before, in the AUTOMATIC scheme of the proposed method, the area (a) and standard deviation (s) atributes are only used in constructing our attribute profiles, since these attributes can be adjusted automatically and are well-related to the object hierarchy in the images. It should be noted that in order to provide a comparative evaluation of the results, the result of the MANUAL scheme and AUTOMATIC scheme are evaluated separately.

The following measures are used in order to evaluate the performance of different classification methods.

1) Average Accuracy (AA): This index shows the average value of the class classification accuracy.

2) Overall Accuracy (OA): This index represents the number of samples which is classified correctly divided by the number of test samples.

3) Kappa Coefficient: This index provides information regarding the amount of agreement corrected by the level of agreement that could be expected due to chance alone.

4) McNemar's Test: This test is used to assess classification results and is calculated by

$$M = \frac{d_{21} - d_{12}}{\sqrt{d_{12} + d_{21}}} \tag{6}$$

where d_{12} is the number of samples that are incorrectly classified by a first classifier but not the second one and d_{21} has a dual meaning [36]. The difference between the proposed method and others is statistically significant at 5-percent significant level if $|M| > 1.96$. It should be noted that in each comparison, the method which provides a better OA has been considered as classifier 1 and the other method as classifier 2.

5) CPU Processing Time: This measure shows the speed of different algorithms. It should be noted, since in all algorithms (except Spectral), EMAP is carried out, the CPU processing time of this step is discarded from all methods. Hence, the CPU processing time is only provided for AP, AP + Spectral, DAFE and NWFE. All methods used were programmed in MATLAB on a computer having Intel(R) Pentium(R) 4 CPU 3.20 GHz and 4GB of memory.

C. Results

I. MANUAL scheme

1. Pavia University

Fig. 4. Classification maps of different methods for Pavia University: a) Spectral b) AP c) Spectral+AP d) DAFE e) NWFE.

TABLE I
PAVIA UNIVERSITY: THE NUMBER OF TRAINING AND TEST SAMPLES; CLASSIFICATION ACCURACIES OF TEST SAMPLES IN PERCENTAGE, THE NUMBER OF FEATURES ARE GIVEN IN BRACKETS.

Class		No. of Samples		Spectral	Attribute Profile	Spectral + AP	DAFE	DBFE	NWFE
No.	Name	Training	Test	(103)	(99)	(103 + 99)	(6 + 8)	(29 + 25)	(11 + 8)
1	Asphalt	548	6631	80.8	**96.4**	95.9	93.7	96.0	93.4
2	Meadows	540	18649	56.1	92.5	94.2	**97.3**	95.3	95.6
3	Gravel	392	2099	53.5	68.2	67.6	**96.2**	75.2	57.6
4	Trees	524	3064	98.7	98.4	**99.8**	97.1	96.8	99.2
5	Metal Sheets	265	1345	99.1	99.5	**99.6**	**99.6**	99.7	99.5
6	Soil	532	5029	78.1	68.4	68.4	**98.3**	88.6	97.7
7	Bitumen	375	1330	84.3	**99.9**	**99.9**	**99.9**	99.9	99.5
8	Bricks	514	3682	91.0	99.5	99.4	**99.5**	99.5	98.8
9	Shadows	231	947	98.3	99.7	99.7	88.4	99.3	**99.6**
AA		–	–	82.25	91.43	91.66	**96.72**	94.55	93.51
OA		–	–	71.64	90.74	90.90	**97.00**	94.55	94.58
Kappa		–	–	0.6511	0.8773	0.8794	**0.9604**	0.9280	0.9287

As can been seen in Table I, DAFE gives the highest classifications accuracy compared to other methods used and improves the overall accuracy of Spectral, AP, Spectral+AP, DBFE and NWFE by almost 25, 6, 6, 2.5 and 2.5 percent, respectively. This shows that when a sufficient set of training samples is available, DAFE leads to more discriminant features in comparison with those achieved by DBFE and NWFE. The main reason for that might be the number of selected features used by DBFE and NWFE is not sufficient. As a result, more features needs to be considered in order to provide more promising results in the case of NWFE and DBFE.

AP shows a better performance than Spectral in terms of accuracies and improves the overall accuracy by almost 19 percent. Spectral has better class accuracies for class Trees and Soils where spectral information can lead to better discrimination of those classes than the spatial information. According to Fig. 4, by considering the spatial dependencies using Attribute Profile, the noisy behavior of classified pixels by RF has been decreased significantly.

As can be seen from the table, by considering both Spectral and AP in the same stacked vector (Spectral + AP), the overall accuracy of the classification is improved by only 0.2% (Table I) the CPU processing time is increased by 63 seconds (Table II). This infers that having both the spectral and AP features in the same vector and using more features (202 features instead of 99 features) does not necessarily lead to better classification results. Spectral+AP improves the class accuracies of Meadows, Trees and Metal Sheets in comparison with the cases when Spectral and AP have been classified separately and degrades the class accuracies of Asphalt, Gravel and Bricks compared with AP and the class accuracy of Soil compared with Spectral. In other words, the consideration of the full features obtained by AP along with the input data (Spectral) sometimes can lead to a better discrimination of different classes and sometimes downgrades class accuracies in comparison with the individual use of either AP or Spectral.

Table II shows that DAFE has the least CPU processing time in comparison with the other methods used. DAFE is a very fast feature extraction method and is able to find more effective features in less CPU processing time than NWFE.

In Table III, the difference in classification accuracy between the DAFE and others are statistically significant using the 5-percent level of significance. In the same way, in comparison with other methods used, NWFE is

TABLE II
PAVIA UNIVERSITY: CPU PROCESSING TIME IN SEC. FOR DIFFERENT METHODS OF THE GENERAL METHOD. SINCE AP ARE USED FOR ALL METHODS, THE CPU PROCESSING TIME FOR MAKING AP IS DISREGARDED. FOR DAFE AND NWFE, THE CPU PROCESSING TIME IS THE SUMMATION OF THE FE PART AND THE CLASSIFICATION STEP.

Spectral	Attribute Profile	Spectral + AP	DAFE	DBFE	NWFE
62	39	102	12	65	50

TABLE III
PAVIA UNIVERSITY: THE RESULT OF MCNEMAR'S TEST TO VALIDATE WHETHER THE DIFFERENCE BETWEEN CLASSIFICATION ACCURACIES OF THE PROPOSED METHOD IS SIGNIFICANTLY DIFFERENT FROM OTHER METHODS.

Pavia University	M
DAFE vs. Spec	99.01
DAFE vs. AP	41.34
DAFE vs. ALL	41.28
DAFE vs. NWFE	20.16
DAFE vs. DBFE	22.70
NWFE vs. ALL	27.61
NWFE vs. AP	27.63
NWFE vs. Spectral	94.55
NWFE vs. DBFE	0.19
AP vs. Spectral	74.75

statistically significant when considering the 5-percent level of significance. As can be seen from Table III, the difference between NWFE and DBFE is not statistically significant at the 5-percent significance level.

2. Pavia Center

In Table IV it can be seen that NWFE works better than DAFE in terms of the classification accuracies. The main reason for that might be 1) the between-scatter matrix in DAFE is not full rank and its rank is equal to L-1 and only L-1 features are selected when using DAFE. As it was mentioned before, since in real situations, the data distribution is complicated, using only L-1 features is not enough. 2) DAFE works well when the distribution of data is normal. Otherwise, the performance of DAFE is not satisfactory. As also can be seen, NWFE improves the result of DAFE by 1 percent in terms of overall accuracy. NWFE improves the overall accuracy of the classification obtained by DBFE. That infers that for Pavia Center, NWFE which is a non-parametric FE can discriminate different classes of interest in comparison with DAFE and DBFE which are parametric and based on a normal distribution.

AP works better than Spectral + AP in terms of overall accuracy and Kappa coefficient. The number of features in Spectral + AP is 201 which increases the possibility of the problems with the curse of dimensionality. In other words, by increasing the dimensionality of the data set, although class separability increases, the accuracy of the class statistics estimation decreases which means a higher dimensional set of statistics must be estimated with a fixed number of samples.

Classification of AP also works better than classification of Spectral in terms of the classification accuracies. One reason for that would be, since this data set contains a very dense urban area, Attribute Profile can provide discriminative information which might be more useful than spectral information.

The CPU processing time for different methods is given in Table V. DAFE has the lowest CPU processing time. The main reasons can be 1) DAFE is a very fast feature extraction approach and 2) only 15 features are classified by RF. Spectral + AP has the worst performance in terms of CPU processing time since 201 features need to be classified by RF.

Fig. 5 shows the classification map for different classifiers. As can be seen, spectral-spatial methods improve the noisy behavior of of using RF on the original data since those methods consider spatial dependencies as well.

As can be seen from Table VI, the difference in classification accuracy between the proposed method using NWFE and the others are statistically significant using the 5-percent level of significance.

Based on the results given in this section, it is obvious that the importance of including the spatial information leads to an increase in classification accuracies when compared to the classification of the original data set. Moreover, considering EMAP allows us to obtain a representation of the image based on complementary characteristics, which helps to a great extend in improving the result of the classification in terms of accuracies. Furthermore, the redundancy of the AP and the original data set can be easily solved by using supervised FE.

II. AUTOMATIC scheme

Table VII shows the result of the classification in terms of the accuracies for the automatic schemes of the proposed method for both test cases. As can be seen from Table VII, the achieved accuracies are almost the same as the accuracies reported in Tables I and IV. However, for the automatic scheme there is no need to adjust the initial parameters for the attribute profiles which is considered as the main shortcoming of the usage of AP. The little difference obtained in the classification accuracies between the MANUAL and AUTOMATIC settings of the proposed method can show that the use of only two attributes; area and standard deviation can model the spatial information on the used data sets considerably and other attributes (diagonal of the box bounding the region and the moment of inertia) do not add significant improvement to classification accuracies although they carry information on the shape of regions. It is generally accepted that the use of different attributes will lead to the extraction of complementary (and redundant) information from the scene leading to increased accuracies when used in classification (provided Hughes effect is efficiently solved by only keeping those features which are most informative).

As can be seen in Table VII, for Pavia University, DAFE works better than the other methods used in the experiment in terms of accuracies and improves the overall accuracy of DBFE and NWFE by almost 2 and 2.5 percent, respectively. It should be noted that, DAFE provides very good accuracies by considering only 14 features. DBFE works slightly better than NWFE in terms of accuracies but with a higher number of features.

TABLE IV
PAVIA CENTER: THE NUMBER OF TRAINING AND TEST SAMPLES; CLASSIFICATION ACCURACIES OF TEST SAMPLES IN PERCENTAGE, THE NUMBER OF FEATURES IN ARE GIVEN BRACKETS.

Class No.	Class Name	No. of Samples Training	No. of Samples Test	Spectral (102)	Attribute Profile (99)	Spectral + AP (102 + 99)	DAFE (7 + 8)	DBFE (30 + 25)	NWFE (13 + 33)
1	Water	824	65147	98.8	98.6	98.2	98.9	96.9	**99.3**
2	Trees	820	6778	88.1	91.0	88.7	88.3	91.2	**92.7**
3	Meadows	824	2266	95.9	95.6	**98.1**	96.3	95.9	95.8
4	Bricks	808	1891	66.4	99.0	99.1	**99.6**	98.8	**99.6**
5	Soil	820	5764	89.9	**99.7**	99.5	98.5	98.4	98.8
6	Asphalt	816	8432	93.7	**99.4**	99.3	99.2	98.6	99.0
7	Bitumen	808	6479	93.3	98.1	98.1	99.4	**99.1**	97.9
8	Tiles	1260	41566	97.6	99.6	99.7	99.7	99.7	**99.8**
9	Shadows	476	2387	99.6	98.5	99.1	63.6	**100**	**100**
AA		–	–	91.51	97.77	97.81	93.71	97.66	**98.14**
OA		–	–	96.56	98.58	98.39	98.05	97.83	**99.02**
Kappa		–	–	0.9503	0.9795	0.9767	0.9717	0.9688	**0.9858**

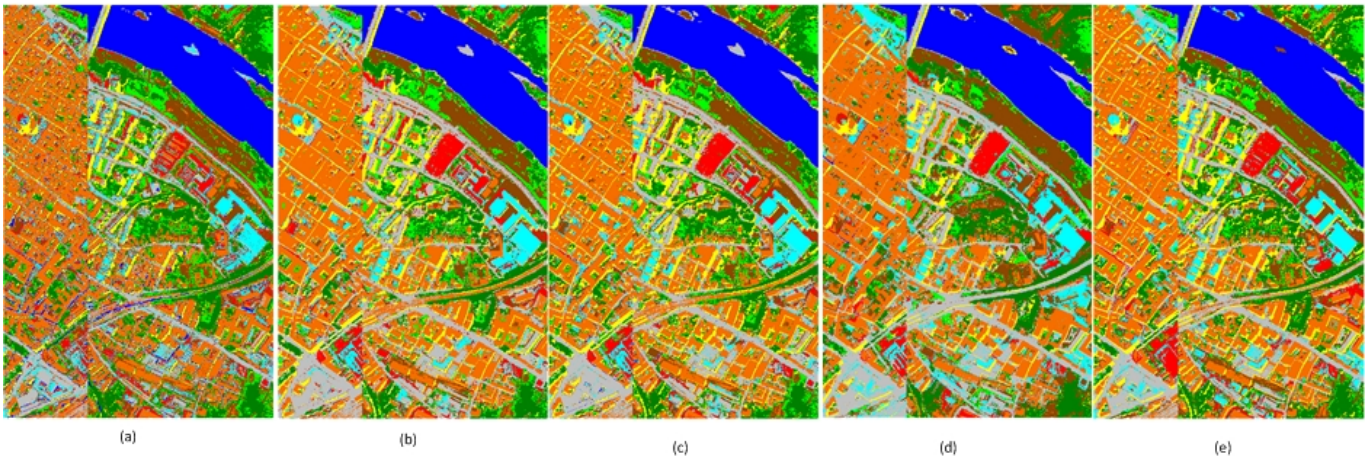

Fig. 5. Classification maps of different methods for Pavia Center: a) Spectral b) AP c) Spectral+AP d) DAFE e) NWFE.

TABLE V
PAVIA CENTER: CPU PROCESSING TIME IN SEC. FOR DIFFERENT METHODS OF THE GENERAL METHOD. SINCE AP ARE USED FOR ALL METHODS, THE CPU PROCESSING TIME FOR MAKING AP IS DISREGARDED. FOR DAFE AND NWFE, THE CPU PROCESSING TIME IS THE SUMMATION OF THE FE PART AND THE CLASSIFICATION STEP.

Spectral	Attribute Profile	Spectral + AP	DAFE	DBFE	NWFE
447	313	1096	32	124	156

TABLE VI
PAVIA CENTER: THE RESULT OF MCNEMAR'S TEST TO VALIDATE WHETHER THE DIFFERENCE BETWEEN CLASSIFICATION ACCURACIES OF THE PROPOSED METHOD IS SIGNIFICANTLY DIFFERENT FROM OTHER METHODS.

Pavia Center	M
NWFE vs. Spec	54.36
NWFE vs. AP	15.48
NWFE vs. ALL	23.63
NWFE vs. DAFE	28.20
NWFE vs. DBFE	34.34
DAFE vs. ALL	-9.28
DAFE vs. AP	-14.90
DAFE vs. Spectral	28.89
DAFE vs. DBFE	26.83
AP vs. Spectral	44.27

For Pavia Center, NWFE works better than others in terms of accuracies. Furthermore, DAFE is more accurate than DBFE with a less number of features. It should be noted that both data sets follow the same trend as both the AUTOMATIC and MANUAL schemes.

Table VIII shows the CPU processing time of the AUTOMATIC scheme of the proposed method in sec. for the University of Pavia and Pavia Center data sets. As expected for both data sets, DAFE has the least CPU processing time. Generally, the AUTOMATIC scheme follows the same trend of the MANUAL scheme and for Pavia University NWFE works better than DBFE and for Pavia Center, DBFE has a shorter processing time than NWFE.

In summary, it can be concluded that the AUTOMATIC version of the proposed method can provide classification maps comparable with the MANUAL version of the proposed method in terms of both classification accuracies and CPU processing time when only two attributes (area and standard deviation) are used instead of four (area, standard deviation, moment of inertia and diagonal of the box). However, the whole procedure in the AUTOMATIC version of the proposed method as the name indicates is automatic and there is no need for any parameters to be set.

Based on our literature review, the proposed method im-

TABLE VII
THE RESULT OF THE CLASSIFICATION IN TERMS OF THE ACCURACIES FOR THE AUTOMATIC SCHEMES FOR THE PROPOSED METHOD FOR BOTH TEST CASES.

Class No.	Pavia Center DAFE (7+6)	Pavia Center DBFE (30+14)	Pavia Center NWFE (13+24)	Pavia Uni DAFE (6+8)	Pavia Uni DBFE (29+16)	Pavia Uni NWFE (11+27)
1	99.3	96.5	**100**	**97.2**	96.9	96.5
2	89.2	86.0	**89.9**	**94.8**	91.4	94.7
3	96.5	96.4	**96.7**	**84.2**	71.6	70.5
4	**99.6**	**99.6**	99.4	99.1	99.8	**99.9**
5	97.9	**99.5**	98.3	99.6	**99.9**	99.6
6	98.1	97.4	**98.7**	**99.1**	98.6	85.0
7	**98.8**	96.9	97.8	99.9	**100**	99.7
8	**99.8**	**99.5**	99.8	**99.8**	99.5	99.5
9	99.6	**100**	**100**	96.4	**99.7**	**99.7**
AA	97.69	96.89	**97.87**	**96.72**	95.30	93.96
OA	98.83	97.19	**99.16**	**96.30**	94.20	93.92
Kappa	0.9830	0.9596	**0.9878**	**0.9514**	0.9244	0.9197

TABLE VIII
PAVIA CENTER & PAVIA UNI: CPU PROCESSING TIME OF THE AUTOMATIC SCHEME OF THE PROPOSED METHOD IN SEC. FOR DIFFERENT METHODS PER SECOND. SINCE AP ARE USED FOR ALL METHODS, THE CPU PROCESSING TIME FOR MAKING AP IS DISREGARDED. FOR DAFE AND NWFE, THE CPU PROCESSING TIME IS THE SUMMATION OF THE FE PART AND THE CLASSIFICATION STEP.

Pavia Center			Pavia Uni		
DAFE	DBFE	NWFE	DAFE	DBFE	NWFE
31	127	149	13	81	63

proves all methods in the literature in terms of classification accuracies. For example, the proposed method improves the classification accuracy of the classification technique proposed in [21] for Pavia University by almost 10 percentage points (the best OA for the Pavia University data set reported in [21] is achieved by DBFE 95% (Table V in [21]) which is equal to 87.97% (same size of train and test sets). The best OA for the Pavia Center data set in [21] is achieved by NFWE and is on the other hand equal to 98.87% (same size of train test set, but slightly larger test set). Thus, the best improvement in OA is from 87.97% to 97.00% (almost by 10 percentage points with the MANUAL version) for Pavia University data set). Based on the results reported in [37], the performance of the proposed method

- against the ICA results in [37], the proposed approach gives improved overall accuracies of 2.5% (with the MANUAL version) and 1.83% (with the AUTOMATIC version) for the Pavia University data set.
- against the PCA results in [37], the proposed approach gives improved overall accuracies of 19.19% (with the MANUAL version) and 18.49% (with the AUTOMATIC version) for Pavia University data set. It should be noted, the test samples of the Pavia Center used in [37], are different from the test samples of Pavia Center used in this work. Therefore, the results reported in [37] and in this paper for Pavia Center are not fully comparable.

It should be again noted that in the proposed AUTOMATIC method, there is no need to adjust any parameters which increases the desirability of using the method.

The following results have been figured out in this work:

- When the number of training samples is adequate, the use of DAFE may lead to better classification accuracies. With reference to Table I, DAFE improves the overall accuracy of NWFE by almost 2.5 percent.
- Fig. 6 shows that not only the number of training sample is important for the efficiency of DAFE and NWFE, but also the distribution of training samples on the whole data set is of importance. As an example, the black boxes in Fig. 6 show two parts of the input data which do not contain training samples. In this case, although the overall accuracy of DAFE (97.00%) is significantly higher than the overall accuracy of NWFE (94.58%), some objects are

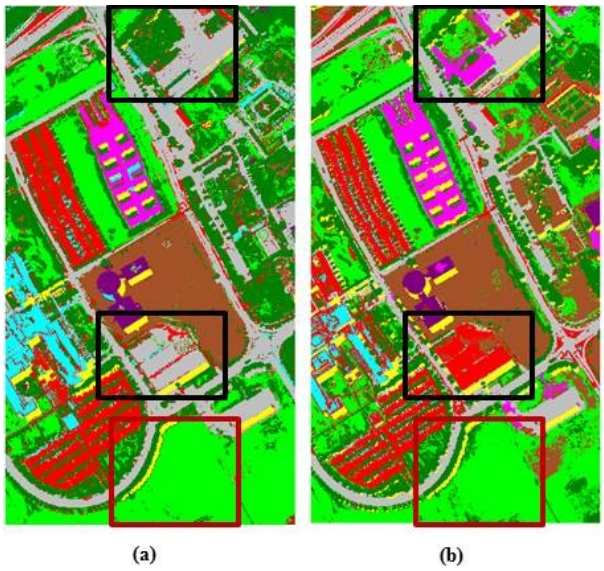

Fig. 6. Comparison between classification maps obtained by (a) DAFE and (b) NWFE

missing in the classification map obtained by DAFE since the data do not have training samples in those regions. On the contrary, for the region where there is an adequate number of training samples (the red box), the use of DAFE leads to a smoother classification map.

- Classification of AP works much better than classification of Spectral in terms of the classification accuracies. In data sets which contain a very dense urban area, AP can provide discriminative information which can be more useful than spectral information.
- The AUTOMATIC version of the proposed method can provide classification maps that are comparable with the MANUAL version of the proposed method in terms of both classification accuracies and CPU processing time when only two attributes (area and standard deviation) are used instead of four (area, standard deviation, moment of inertia and diagonal of the box). However, the advantage of the AUTOMATIC version is that there is no need for any parameters to be set.

IV. CONCLUSION

In this paper, a new approach is proposed for the classification of hyperspectral images which uses both spectral and spatial information. The method can be implemented fully automatically. In order to use the spatial information, attribute profiles are taken into account. For reducing the redundancy of both the spatial information and the original spectral data in order to provide more accurate classification results, a few supervised feature extraction methods are considered. The new method was tested on two data sets, and the obtained results confirm that considering spatial information by using attribute profiles in conjunction with spectral information can significantly improve the classification accuracies of the original data. In addition, by using supervised feature extraction, the classification accuracies can be increased further. Furthermore, in order to avoid the main difficulties of using attribute profiles, an automatic version of the proposed method is introduced which only considers area and standard deviation attributes. The AUTOMATIC method obtained almost the same results as the MANUAL method in terms of the classification accuracies and CPU processing time and solved the main difficulty of the MANUAL method which is related to the initialization of the parameters in the EMAP. The proposed method was tested on two widely used ROSIS data sets named Pavia Center and Pavia University. The proposed method worked well in terms of the classification accuracy and CPU processing time which confirms the ability of the method to classify high dimensional data sets.

In experiments, the proposed approach can be thought of as a general framework and some of its steps can be replaced by other techniques, possibly to improve the CPU processing time and classification accuracies for the proposed approach. For example, other types of feature reduction techniques such as Kernel PCA, ICA and supervised feature extraction/selection can be used instead of PCA in the proposed approach.

V. ACKNOWLEDGMENT

The ROSIS data and corresponding reference information were obtained from Prof. Paolo Gamba of the University of Pavia, Italy. This work was supported in part by Icelandic Research Fund for Graduate Students.

REFERENCES

[1] M. Fauvel, Y. Tarabalka, J. A. Benediktsson, J. Chanussot, and J. C. Tilton, "Advances in spectral-spatial classification of hyperspectral images," *Proceedings of the IEEE*, vol. 101, no. 3, pp. 652–675, 2013.

[2] G. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. Inf. Theory*, vol. IT, no. 14, pp. 55 – 63, 1968.

[3] D. A. Landgrebe, *Signal Theory Methods in Multispectral Remote Sensing*. Hoboken, NJ: Wiley, 2003.

[4] P. Ghamisi, M. S. Couceiro, and J. A. Benediktsson, "Classification of hyperspectral images with binary fractional order darwinian pso and random forests," pp. 88 920S–88 920S–8, 2013. [Online]. Available: http://dx.doi.org/10.1117/12.2027641

[5] S. Tadjudin and D. Landgrebe, "Classification of high dimensional data with limited training samples," in *Tech. Rep., School of Electrical and Computer Engineering, Purdue University*, 1998.

[6] H. Derin and P. A. Kelly, "Discrete-index Markov-type random processes," *Proceedings of the IEEE*, vol. 77, no. 10, pp. 1485–1510, 1989.

[7] F. Bovolo and L. Bruzzone, "A context-sensitive technique based on support vector machines for image classification," *Proc. PReMI*, pp. 260–265, 2005.

[8] D. Liu, M. Kelly, and P. Gong, "A spatial-temporal approach to monitoring forest disease spread using multi-temporal high spatial resolution imagery," *Remote Sens. Environment*, vol. 101, no. 10, p. 167180, 2006.

[9] G. Moser, S. B. Serpico, and J. A. Benediktsson, "Land-cover mapping by Markov modeling of spatial-contextual information in very-high-resolution remote sensing images," *Proceedings of the IEEE*, vol. 101, March 2013.

[10] G. Moser and S. B. Serpico, "Combining support vector machines and Markov random fields in an integrated framework for contextual image classification," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. PP, no. 99, pp. 1–19, 2012.

[11] M. Khodadadzadeh, R. Rajabi, and H. Ghassemian, "Combination of region-based and pixel-based hyperspectral image classification using erosion technique and MRF model," *Electrical Engineering (ICEE), 2010 18th Iranian Conference on*, pp. 294 –299, may 2010.

[12] P. Ghamisi, J. A. Benediktsson, and M. O. Ulfarsson, "Spectral-spatial classification of hyperspectral images based on hidden Markov random fields," *IEEE Trans. on Geos. and Remote Sens.*, vol. PP, no. 99, pp. 1–10, 2013. [Online]. Available: http://dx.doi.org/10.1109/TGRS.2013.2263282

[13] G. Zhang, X. Jia, and N. M. Kwok, "Spectral-spatial based super pixel remote sensing image classification," *Proc. 4th Int. Congr. Image Signal Process.*, no. 3, p. 16801684, 2011.

[14] P. Ghamisi, M. S. Couceiro, N. M. F. Ferreira, and L. Kumar, "Use of darwinian particle swarm optimization technique for the segmentation of remote sensing images," *IGARSS*, 2012.

[15] P. Ghamisi, M. S. Couceiro, J. A. Benediktsson, and N. M. F. Ferreira, "An efficient method for segmentation of images based on fractional calculus and natural selection," *Expert Syst. Appl.*, vol. 39, no. 16, pp. 12 407–12 417, 2012.

[16] Y. Tarabalka, J. A. Benediktsson, and J. Chanussot, "Spectral-spatial classification of hyperspectral imagery based on partitional clustering techniques,," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 47, no. 5, pp. 2973–2987, 2009.

[17] P. Ghamisi, M. S. Couceiro, F. M. Martins, and J. A. Benediktsson, "Multilevel image segmentation approach for remote sensing images based on fractional-order darwinian particle swarm optimization," *IEEE Trans. on of Remote Sens. and Geos.*, In press.

[18] P. Ghamisi, M. Couceiro, M. Fauvel, and J. A. Benediktsson, "Integration of segmentation techniques for classification of hyperspectral images," *IEEE Geos. and Remote Sens. Letter*, In Press.

[19] M. Pesaresi and J. A. Benediktsson, "A new approach for the morphological segmentation of high-resolution satellite imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 2, pp. 309–320, 2001.

[20] J. A. Palmason, J. A. Benediktsson, J. R. Sveinsson, , and J. Chanussot, "Classication of hyperspectral data from urban areas using morphological preprocessing and independent component analysis," *IEEE International Geoscience and Remote Sensing Symposium (IGARSS05)*, no. 3, p. 176 179, 2005.
[21] M. Fauvel, J. A. Benediktsson, J. Chanussot, and J. R. Sveinsson, "Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles," *IEEE Trans. Geos. and Remote Sens.*, vol. 46, no. 11, pp. 3804–3814, 2008.
[22] M. D. Mura, J. A. Benediktsson, B. Waske, and L. Bruzzone, "Morphological attribute profiles for the analysis of very high resolution images," *IEEE Trans. Geos. and Remote Sens.*, vol. 48, no. 10, pp. 3747–3762, 2010.
[23] M. Pedergnana, P. R. Marpu, M. D. Mura, J. A. Benediktsson, and L. Bruzzone, "A novel technique for optimal feature selection in attribute profiles based on genetic algorithms," *IEEE Trans. Geo. and Remote Sens.*,, vol. PP, no. 99, p. 1 15, 2005.
[24] I. Jolliffe, *Principal Component Analysis*, ser. Springer Series in Statistics. Springer, 2002. [Online]. Available: http://books.google.is/books?id=_olByCrhjwIC
[25] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. Academic Press, 1974.
[26] L. Jimenez and D. A. Landgrebe, "Hyperspectral data analysis and supervised feature reduction via projection pursuit," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 37, no. 6, pp. 2653–2667, 1999.
[27] C. Lee and D. A. Landgrebe, "Feature extraction based on decision boundaries," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 4, pp. 388–400, 1993.
[28] B. C. Kuo and D. A. Landgrebe, "A robust classication procedure based on mixture classiers and nonparametric weighted feature extraction," *Comput.Vis. ImageUnderst.*, vol. 64, no. 3, p. 377389, 1996.
[29] E. J. Breen and R. Jones, "Attribute openings, thinnings and granulometries," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 11, pp. 2486–2494, 2013.
[30] N. Bouaynaya and D. Schonfeld, "Theoretical foundations of spatially-variant mathematical morphology part ii: Gray-level images," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 30, no. 5, pp. 837–850, 2008.
[31] M. D. Mura, J. A. 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.
[32] P. Salembier, A. Oliveras, , and L. Garrido, "Anti-extensive connected operators for image and sequence processing," *IEEE Trans. Image Process.*, vol. 7, no. 4, pp. 555–570, 1998.
[33] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, p. 532, 2001.
[34] ——, "RF tools a class of two eyed algorithms," *SIAM Workshop*, 2003.
[35] P. Marpu, M. Pedergnana, M. D. Mura, J. A. Benediktsson, and L. Bruzzone, "Automatic generation of standard deviation attribute profiles for spectral -spatial classification of remote sensing data," *IEEE Geoscience and Remote Sensing Letters*, vol. 10, no. 2, pp. 293–297, 2013.
[36] G. M. Foody, "Thematic map comparison: Evaluating the statistical significance of differences in classification accuracy," *PERS*, vol. 70, no. 5, pp. 627–633, 2004.
[37] M. D. Mura, A. Villa, J. A. Benediktsson, J. Chanussot, and L. Bruzzone, "Classification of hyperspectral images by using extended morphological attribute profiles and independent component analysis," *IEEE Geoscience and Remote Sensing Letters*, vol. 8, no. 3, 2011.

Pedram Ghamisi graduated with a B.Sc. degree in Civil (Survey) Engineering from the Tehran South Campus of Azad University. Then, he obtained the M.Sc. degree in Remote Sensing at K.N.Toosi University of Technology in 2012. He received the Best Researcher Award for M.Sc. students in K. N. Toosi University of Technology in the academic year 2010-2011. Mr. Ghamisi was the recipient of the IEEE Mikio Takagi Prize which was awarded for the first place in the Student Paper Competition at the 2013 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Melbourne, July 2013. He is currently a Ph.D. student in Electrical and Computer Engineering at the University of Iceland. His research interests are in remote sensing and image analysis with the current focus on spectral and spatial techniques for hyperspectral image classification. He serves as a reviewer for a number of journals including IEEE Trans. Image Processing, IEEE JSTARS and IEEE GRSL.

Jón Atli Benediktsson Jón Atli Benediktsson received the Cand.Sci. degree in electrical engineering from the University of Iceland, Reykjavik, in 1984, and the M.S.E.E. and Ph.D. degrees from Purdue University, West Lafayette, IN, in 1987 and 1990, respectively. He is currently Pro Rector for Academic Affairs and Professor of Electrical and Computer Engineering at the University of Iceland. His research interests are in remote sensing, biomedical analysis of signals, pattern recognition, image processing, and signal processing, and he has published extensively in those fields. Prof. Benediktsson was the 2011-2012 President of the IEEE Geoscience and and Remote Sensing Society (GRSS) and has been on the GRSS AdCom since 2000. He was Editor of the IEEE Transactions on Geoscience and Remote Sensing (TGRS) from 2003 to 2008 and has served as Associate Editor of TGRS since 1999, the IEEE Geoscience and Remote Sensing Letters since 2003 and IEEE Access since 2013. He is on the International Editorial Board of the International Journal of Image and Data Fusion and was the Chairman of the Steering Committee of IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (J-STARS) 2007-2010. Prof. Benediktsson is a co-founder of the biomedical start up company Oxymap (www.oxymap.com). He is a Fellow of the IEEE and a Fellow of SPIE. He received the Stevan J. Kristof Award from Purdue University in 1991 as outstanding graduate student in remote sensing. In 1997, Dr. Benediktsson was the recipient of the Icelandic Research Council's Outstanding Young Researcher Award, in 2000, he was granted the IEEE Third Millennium Medal, in 2004, he was a co-recipient of the University of Iceland's Technology Innovation Award, in 2006 he received the yearly research award from the Engineering Research Institute of the University of Iceland, and in 2007, he received the Outstanding Service Award from the IEEE Geoscience and Remote Sensing Society. He is co-recipient of the 2012 IEEE Transactions on Geoscience and Remote Sensing Paper Award and the co-recipient of the 2013 IEEE Geoscience and Remote Sensing Society Highest Impact Paper Award. He is a member of the Association of Chartered Engineers in Iceland (VFI), Societas Scinetiarum Islandica and Tau Beta Pi.

Johannes R. Sveinsson Johannes R. Sveinsson received the B.S. degree from the University of Iceland, Reykjavk, and the M.S. and Ph.D. degrees from Queens University, Kingston, ON, Canada, all in electrical engineering. He is currently the Head and Professor with the Department of Electrical and Computer Engineering, University of Iceland, he was with the Laboratory of Information Technology and Signal Processing from 1981 to 1982 and, from November 1991 to 1998, the Engineering Research Institute and the Department of Electrical and Computer Engineering as a Senior Member of the research staff and a Lecturer, respectively. He was a Visiting Research Student with the Imperial College of Science and Technology, London, U.K., from 1985 to 1986. At Queens University, he held teaching and research assistantships. His current research interests are in systems and signal theory. Dr. Sveinsson received the Queens Graduate Awards from Queens University. He was a co-recipient of the 2013 IEEE GRSS Highest Impact Paper Award.