Chapter 12

Decision Fusion for Hyperspectral Classification

Mathieu Fauvel[*◊], Jocelyn Chanussot[*] and Jon Atli Benediktsson[◊]

[*]Laboratoire des Images et des Signaux - LIS-GIPSA/INPG
BP 46 - 38402 Saint Martin d'Hères - FRANCE
E-mail: {*mathieu.fauvel, jocelyn.chanussot*}*@lis.inpg.fr*

[◊]Department of Electrical and Computer Engineering
University of Iceland, Hjardarhagi 2-6, 107 Reykjavik, ICELAND
E-mail: {*benedikt*}*@hi.is*

Abstract

In the recent years, pixel-wise classification of hyperspectral images aroused many developments, and the literature now provides various classifiers for numerous applications. In this chapter, we present a generic framework where the redundant or complementary results provided by multiple classifiers can actually be aggregated. Taking advantage from the specificities of each classifier, the decision fusion thus increases the overall classification performances. The proposed fusion approach is in two steps. In a first step, data are processed by each classifier separately and the algorithms provide for each pixel membership degrees for the considered classes. Then, in a second step, a fuzzy decision rule is used to aggregate the results provided by the algorithms according to the classifiers' capabilities. The general framework proposed for combining information from several individual classifiers in multiclass classification is based on the definition of two measures of accuracy. The first one is a point-wise measure which estimates for each pixel the reliability of the information provided by each classifier. By modeling the output of a classifier as a fuzzy set, this point-wise reliability is defined as the degree of uncertainty of the fuzzy set. The second measure estimates the global accuracy of each classifier. It is defined *a priori* by the user. Finally, the results are aggregated with an adaptive fuzzy fusion ruled by these two accuracy measures. The method is illustrated by considering the classification of hyperspectral remote sensing images from urban areas. It is tested and validated with two classifiers on a ROSIS image from Pavia, Italy. The proposed method improves the classification results when compared with the separate use of the different classifiers. The approach is also compared to several other standard fuzzy fusion schemes.

Keywords: Data fusion, classification, remote sensing, fuzzy logic, fuzzy set theory, decision fusion.

Contents

12 Decision Fusion for Hyperspectral Classification **1**
12.1 Introduction . . . 1
12.2 Fuzzy set theory . . . 3
12.2.1 Fuzzy set theory . . . 3
12.2.2 Class representation . . . 6
12.3 Information Fusion . . . 7
12.3.1 Introduction . . . 7
12.3.2 Measure of confidence . . . 9
12.3.3 Combination operator . . . 10
12.4 The Fusion scheme . . . 11
12.5 Experimental Results . . . 11
12.5.1 Test image . . . 11
12.5.2 Classifier based on mathematical morphology and neural network . . . 12
12.5.3 Classifier based on Support Vector Machines . . . 16
12.5.4 Decision fusion . . . 20
12.6 Conclusion . . . 21

12.1 Introduction

With the development of remote sensing sensors, hyperspectral remote sensing images are now widely available. They are characterized by hundreds of spectral bands. For a classification task, the increased dimensionality of the data increases the capability to detect various classes with a better accuracy. But at the same time, classical classification techniques are facing the problem of statistical estimation in high dimensional space. Due to the high number of features and small number of training samples, reliable estimation of statistical parameters is difficult [1]. Furthermore, it is proved that, with a limited training set, beyond a certain limit, the classification accuracy decreases as the number of features increases (Hughes phenomenon [2]). However, several classification algorithms have been proposed in the past few years.

Recently, *support vector machines* (SVMs) have shown to be well suited for high dimensional classification problems [3,4]. With SVMs, classes are not characterized by statistical criteria but by a geometrical criterion. SVMs seek a separating hyperplane maximizing the distance with the closest training samples for two classes. This approach allows SVMs to have a very high capability of generalization and, as a consequence, only require a few training samples. In addition, for non linearly separable data, SVMs use the kernel trick to map the data onto a higher dimensional space where they are linearly separable [5]. Early work in classification of remotely sensed images by SVMs showed promising results [6, 7]. In [8], several SVM-based classifiers are compared with other classical classifiers such as a K-nearest neighbors classifier and a neural network classifier. The SVMs using the kernel trick outperformed the other classifiers in terms of accuracy. Multiclass SVMs performances were also positively compared with a discriminant

analysis classifier, a decision tree classifier and a feedforward neural network classifier with a limited training set [9]. Though these experiments highlight the good generalization capability of SVMs, the data used were pre-processed, *i.e.* 3 selected bands were used for the classification and thus, performances in high dimensional space were not investigated. In both articles [8, 9], the *Gaussian radial basis* kernels were shown to produce the best results. In [10], several spectral-based kernels were tested on hyperspectral data. These kernels were designed to handle spectral meaning, and, in particular, various non-Euclidean metrics were considered to characterize similarity between vectors. In [11], two kernels are considered and compared to assess the generalization capability of SVMs as well as the ability of SVMs to deal with high dimensional feature spaces in the situation of very limited training set.

Another strategy consists in performing a classification with some feature extraction based on Mathematical morphology [12, 13]. Such an approach was initially designed to classify panchromatic data from urban areas: Pesaresi and Benediktsson [14] proposed to construct a morphological profile by a composition of geodesic opening and closing operations with increasing sizes. A neural network approach was used for the pixel-wised classification of the extracted profile. The profile consists of multiple opening and closing transformations of the same base image and should be more effective in discrimination of different urban features. On the other hand, since the images are all transformation of the same image, there may be a lot of redundancy evident in the feature set. Therefore, feature extraction can be desirable in finding the most important features in the feature space. In [15], the method in [14] is extended by including decision boundary feature extraction (DBFE) to reduce the redundancy in the morphological profile. Regarding the extension of this method to hyperspectral images, a simple approach was suggested in [16], consisting in using only the first principal component (PC) of hyperspectral image data to build a morphological profile. In [17], this method is extended. This issue is also addressed by Plaza in [18].

All these methods have their own characteristics and advantages; none of them strictly outperforming all the others.

Usually, for a given data set, performances in terms of *global* and *by class* classification accuracies depend on the considered classes, *i.e.*, on their spectral and spatial characteristics. For instance, methods based on morphological filtering are well suited to classify structures with a typical spatial shape, like man-made constructions. On the contrary, algorithms based on statistical approaches perform better for the classification of vegetation and soils. As a consequence, we propose to use several approaches and try to take advantage of the strengths of each algorithm. This concept is called *decision fusion* [19]. Decision fusion can be defined as the process of fusing information from several individual data sources after each data source has undergone a preliminary classification. For instance, Benediktsson *et al.* [19] proposed a multisource classifier based on a combination of several neural/statistical classifiers. The samples are first classified by two classifiers (a neural network and a multisource classifier); every sample with agreeing results is assigned to the corresponding class. In case of conflicting situation, a second neural network is used to classify the remaining samples. The main limitation of this method is the need of large training sets to train the different classifiers. In [20], Jeon *et al.* used two decision fusion rules to classify multitemporal Thematic Mapper data. Recently, Lisini *et al.* proposed to combine sources according to their class accuracies [21]. In this chapter, the decision fusion rule is modeled with fuzzy data fusion rules.

Data fusion has been used successfully in many classification problems. Tupin *et al.* [22] combined several structure detectors to classify SAR images using Dempster-Shafer theory. In [23], Tupin and Roux aggregated information extracted from SAR and optical images for building detection. In a first step, potential buildings edge were extracted from the SAR image and, in a second step, the buildings shape was extracted from the optical image using the SAR image information. Chanussot *et al.* proposed several strategies to combine the output of a line detector applied to multitemporal images [24]. Here, we propose a general framework to aggregate the results of different classifiers. Conflicting situations, where the different classifiers disagree, are solved by estimating the point-wise accuracy and modeling the global reliability for each algorithm [25]. This leads to the definition of an adaptive fusion scheme ruled by these reliability measures. The proposed algorithm is based on fuzzy sets and possibility theory. The framework

of the addressed problem is modeled as follows: for a given data set, n classes are considered, and m classifiers are assumed to be available. For an individual pixel, each algorithm provides as an output a membership degree for each of the considered classes. The set of these membership values is then modeled as a fuzzy set and the corresponding degree of fuzziness determines the local reliability of the algorithm. The global accuracy is manually defined for each class after a statistical study of the results obtained with each separately used classifier. Hence, the fusion is performed by aggregating the different fuzzy sets provided by the different classifiers. It is adaptively ruled by the reliability information and does not require any further training. The decision is postponed to the end of the fusion process in order to take advantage of each algorithm and enable more accurate results in conflicting situations. In previous studies, this method has been successfully applied to panchromatic data. Taking material from [11], this chapter presents the general framework and applies it to the classification of hyperspectral data, using the two previously mentioned algorithms as information sources.

This chapter is organized as follows. Fuzzy set theory and measures of fuzziness are briefly presented in Section 12.2. Section 12.2.2 presents the model for each classifier's output in terms of a fuzzy set. Then, the problem of information fusion is discussed in Section 12.3. The proposed fusion scheme is detailed in Section 12.4 and experimental results are presented in Section 12.5. Finally, conclusions are drawn.

12.2 Fuzzy set theory

Traditional mathematics assigns a membership value of 1 to elements which are members of a set, and 0 to those which are not, thus defining *crisp sets*. On the contrary, *fuzzy set* theory handles the concept of partial membership to a set, with real-valued membership degrees ranging from 0 to 1. Fuzzy set theory was introduced in 1965 by Zadeh as a mean to model the vagueness and ambiguity in complex systems [26]. It is now widely used to process unprecise or uncertain data [27,28]. In particular, it is an appropriate framework to handle the output of one given classifier for further processing since it usually does not come in a binary form and includes some ambiguity. In this section, we first recall general definitions and properties of fuzzy sets. Then, we precise the model used for the representation of the classifiers output.

12.2.1 Fuzzy set theory

Definitions

Definition 1 (Fuzzy subset) *A fuzzy subset*[1] F *of a reference set* U *is a set of ordered pairs* $F = \{(x, \mu_F(x)) \mid x \in U\}$, *where* $\mu_F : U \to [0,1]$ *is the membership function of* F *in* U.

Definition 2 (Normality) *A fuzzy set is said to be* normal *if and only if :* $\max \mu_F(x) = 1$.

Definition 3 (Support) *The support of a fuzzy set* F *is defined as:*

$$Supp(F) = \{x \in U \mid \mu_F(x) > 0\}.$$

Definition 4 (Core) *The core of a fuzzy set is the (crisp) set containing the points with the largest membership value (1). It is empty if the set is non normal.*

Logical operations

Classical Boolean operations extend to fuzzy sets [26]. With F and G two fuzzy sets, classical extensions are defined as follows:

[1] For convenience, we will use the term "fuzzy set" instead of "fuzzy subset" in the following, where a fuzzy set F is described by its membership function μ_F.

Equality The equality between two fuzzy sets is defined as the equality of their membership functions:

$$\mu_F = \mu_G \Leftrightarrow \forall x \in U, \mu_F(x) = \mu_G(x) \tag{12.1}$$

Inclusion The inclusion of a set in another one is defined by the inequality of their membership functions:

$$\mu_F \subset \mu_G \Leftrightarrow \forall x \in U, \mu_F(x) \leq \mu_G(x) \tag{12.2}$$

Union The union of two fuzzy sets is defined by the maximum of their membership functions:

$$\forall x \in U, (\mu_F \cup \mu_G)(x) = \max\{\mu_F(x), \mu_G(x)\} \tag{12.3}$$

Intersection The intersection of two fuzzy sets is defined by the minimum of their membership functions:

$$\forall x \in U, (\mu_F \cap \mu_G)(x) = \min\{\mu_F(x), \mu_G(x)\} \tag{12.4}$$

Complement The complement of a fuzzy set F is defined by:

$$\forall x \in U, \mu_{\overline{F}}(x) = 1 - \mu_F(x) \tag{12.5}$$

Measures of fuzziness

Fuzziness is an intrinsic property of fuzzy sets. To measure to what degree a set is fuzzy, and thus estimate the corresponding ambiguity, several definitions have been proposed [29] [30]. Ebanks [31] proposed to define the degree of fuzziness as a function f with the following properties:

1. $\forall F \subset U$, if $f(\mu_F) = 0$ then F is a crisp set
2. $f(\mu_F)$ is maximum if and only if $\forall x \in U, \mu_F(x) = 0.5$
3. $\forall (\mu_F, \mu_G) \in U^2$, $f(\mu_F) \geq f(\mu_G)$ if $\forall x \in U \begin{cases} \mu_G(x) \geq \mu_F(x) & if \quad \mu_F(x) \geq 0.5 \\ \mu_G(x) \leq \mu_F(x) & if \quad \mu_F(x) \leq 0.5 \end{cases}$
4. $\forall F \in U,\ f(\mu_F) = f(\mu_{\overline{F}})$. A set and its complement have the same degree of fuzziness
5. $\forall (\mu_F, \mu_G) \in U^2, f(\mu_F \cup \mu_G) + f(\mu_F \cap \mu_G) = f(\mu_F) + f(\mu_G)$

Derived from the probability theory and the classical Shannon entropy, De Luca and Termini [29] defined a fuzzy entropy satisfying the above five properties:

$$H_{DTE}(\mu_F) = -K \sum_{i=1}^{n} \Big(\mu_F(x_i) \log_2(\mu_F(x_i)) + (1 - \mu_F(x_i)) \log_2(1 - \mu_F(x_i)) \Big) \tag{12.6}$$

Bezdeck [32] proposed an alternative measure of fuzziness based on a multiplicative class.

Definition 5 (Multiplicative Class) *A multiplicative class is defined as:*

$$H_*(\mu_F) = K \sum_{i=1}^{n} g(\mu_F(x_i)),\ K \in R^+ \tag{12.7}$$

where $g(\mu_F)$ is defined as:

$$\begin{cases} g(t) &= \widetilde{g}(t) - \min\limits_{0 \leq t \leq 1} \widetilde{g}(t) \\ \widetilde{g}(t) &= h(t)h(1-t) \end{cases} \tag{12.8}$$

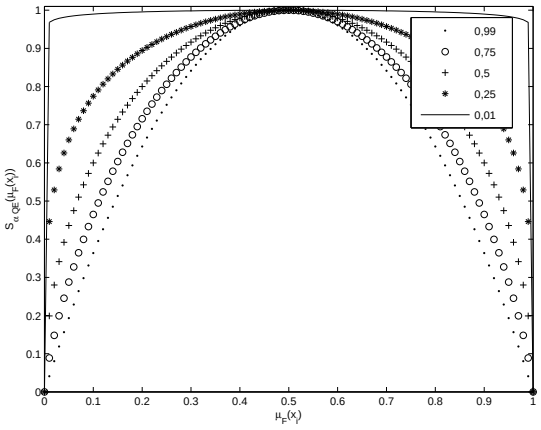


Figure 12.1: Influence of parameter α on $S_{\alpha QE}$

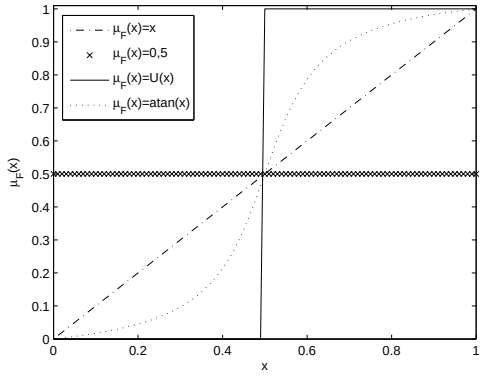


Figure 12.2: Example of four fuzzy sets with different degrees of fuzziness.

and h is a concave increasing function on $[0,1]$:

$$h : [0,1] \rightarrow R^2, \forall x \in [0,1]\ h'(x) > 0\ and\ h''(x) < 0 \tag{12.9}$$

The multiplicative class allows the definition of various fuzziness measures, where different choices of g lead to different behaviors. For instance, let $h : [0,1] \rightarrow R^+$ be $h(t) = t^\alpha,\ 0 < \alpha < 1$. The function h satisfies the required conditions for the multiplicative class, and the function:

$$H_{\alpha QE}(\mu_F) = \frac{1}{n2^{-2\alpha}} \sum_{i=1}^{n} \mu_F(x_i)^\alpha (1 - \mu_F(x_i))^\alpha \tag{12.10}$$

is a measure of fuzziness, namely the $\alpha - Quadratic\ entropy$. Rewriting (12.10) as:

$$\begin{cases} H_{\alpha QE}(\mu_F) &= \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} S_{\alpha QE}(\mu_F(x_i)) \\ S_{\alpha QE}(\mu_F(x_i)) &= \dfrac{\mu_F(x_i)^\alpha (1 - \mu_F(x_i))^\alpha}{2^{-2\alpha}} \end{cases} \tag{12.11}$$

we can analyze the influence of parameter α (see Fig. 12.1): the measure becomes more and more selective as α increases from 0 to 1. With α close to 0, all the fuzzy sets have approximately the same degree of fuzziness and the measure is not sensitive to changes in μ_F, whereas with α close to 1, the measure is highly selective with the degree of fuzziness quickly decreasing when the fuzzy set differs from $\mu_F = 0.5$. As a consequence, an intermediate value such as $\alpha = 0.5$ usually provides a good trade-off [32].

Example of fuzzy sets and their fuzziness values are given in Fig. 12.2 and Table 12.1, respectively. For the binary set, fuzziness is null with respect to the property 1) above. Property 2) is fulfilled since the fuzziness is maximum

Table 12.1: Degree of fuzziness for different fuzzy sets computed with $\alpha - Quadratic\ entropy$

$\alpha =$	0.01	0.25	0.5	0.75	0.99
$H_{\alpha QE}(\mu_F(x)) = atan(x)$	0.959	0.600	0.394	0.271	0.196
$H_{\alpha QE}(\mu_F(x)) = x$	0.967	0.725	0.549	0.423	0.333
$H_{\alpha QE}(\mu_F(x)) = U(x)$	0	0	0	0	0
$H_{\alpha QE}(\mu_F(x)) = 0.5$	0.993	0.840	0.707	0.594	0.503

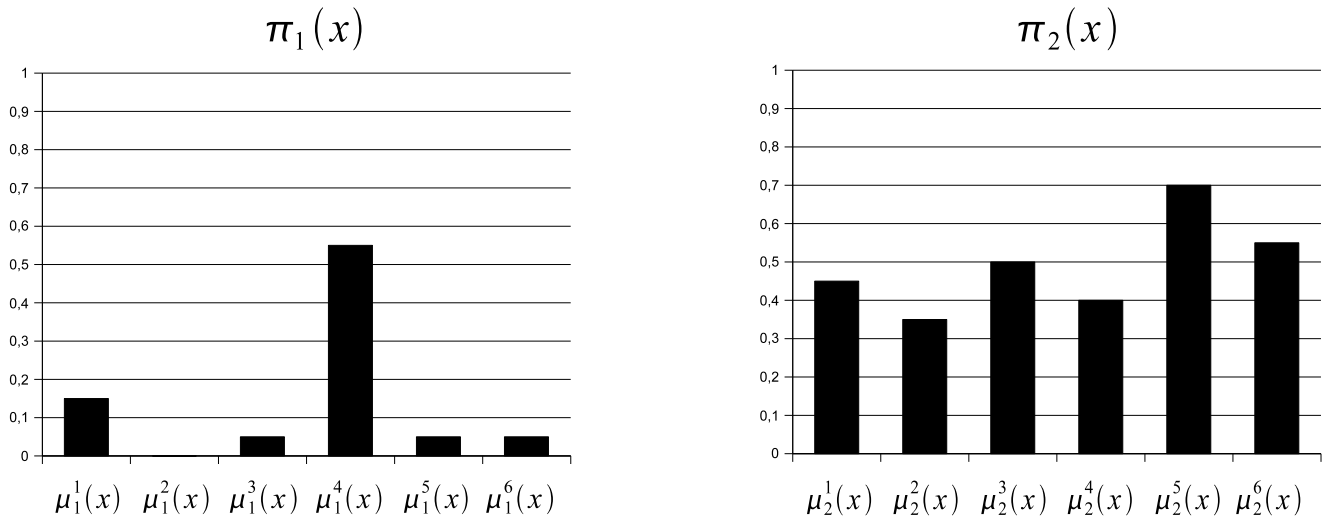


Figure 12.3: Example of two conflicting sets π for a given pixel x

for $\mu_F(x) = 0.5$. Following the condition 3), the fuzzy set with the arctan membership function has a lower fuzziness than the fuzzy set with the linear membership function.

12.2.2 Class representation

An n-classes classification problem is considered, for which m different classifiers are available. For a given pixel x, the output of classifier i is the set of numerical values:

$$\{\mu_i^1(x), \mu_i^2(x), \ldots, \mu_i^j(x), \ldots, \mu_i^n(x)\} \tag{12.12}$$

where $\mu_i^j(x) \in [0,1]$ (after a normalization, if required) is the membership degree of pixel x to class j according to classifier i. The higher this value, the more likely it is that the pixel belongs to class j (if one single classifier is used, the decision is taken by selecting the class j maximizing $\mu_i^j(x)$: $class_{selected}(x) = argmax_j(\mu_i^j(x))$). Depending on the classifier, $\mu_i^j(x)$ can be of different nature: probability, posterior probability at the output of a neural network, membership degree at the output of a fuzzy classifier, *etc.* In any case, the set $\pi_i(x) = \{\mu_i^j(x), j = 1, ..., n\}$ provided by every classifier i can be considered as a fuzzy set.

As a conclusion, for every pixel x, m fuzzy sets are computed, one by each classifier. This set of fuzzy sets constitutes the input for the fusion process:

$$\{\pi_1(x), \pi_2(x), \ldots, \pi_i(x), \ldots, \pi_m(x)\} \tag{12.13}$$

In Fig. 12.3, two conflicting sets are represented: for this pixel, trusting the first classifier (on the left), one would select class number 4, whereas if one trusts the second classifier (on the right), class number 5 would be selected. The handling of such conflicting situations is the central issue that needs to be addressed by the fusion system. As a matter

of fact, the fusion of the non conflicting results is of little interest in our case: though it might increase our believe in the corresponding result, it certainly won't change the final decision, and thus won't increase the classification performances. On the contrary, in the case of conflicting results, at least one classifier is wrong and the fusion gives a chance to correct this and increase the classification performances. Fuzzy set theory provides various combination operators to aggregate these fuzzy sets. Such combination operators are discussed in the next section.

12.3 Information Fusion

After briefly recalling the basics of data fusion, we discuss in this Section the problem of measuring the confidence of individual classifiers. We finally propose an adaptive fusion operator. In the following we denote the fuzzy set i by π_i and m the number of sources.

12.3.1 Introduction

Data fusion consists of combining information from several sources in order to improve the decision [33]. As previously mentioned, the most challenging issue is to solve conflicting situations where some of the sources disagree. Numerous combination operators have been proposed in the literature. They can be classified into three different kinds, depending on their behavior [34]:

- *Conjunctive combination*: This corresponds to a *severe* behavior. The resulting fuzzy set is necessarily smaller than the initial sets and the core is included in the initial cores (it can only decrease). The largest conjunctive operator is the fuzzy intersection (12.4) leading to the following fuzzy set: $\pi_\wedge(x) = \bigcap_{i=1}^{N} \pi_i(x)$. *T-norms* are conjunctive operators. They are commutative, associative, increasing and with $\pi_i(x) = 1$ as a neutral element (*i.e.*, if $\pi_2(x) = 1$ then $\pi_\wedge(x) = \pi_1(x) \cap \pi_2(x) = \pi_1(x)$). They satisfy the following property:

$$\pi_\wedge(x) \leq \min_{i \in [1,m]} \pi_i(x) \tag{12.14}$$

- *Disjunctive combination*: This corresponds to an *indulgent* behavior. The resulting fuzzy set is necessarily larger than the initial sets and the core contains the initial cores (it can only increase). The smallest disjunctive operator is the fuzzy union (12.3), leading to the following fuzzy set: $\pi_\vee(x) = \bigcup_{i=1}^{N} \pi_i(x)$. *T-conorms* are disjunctive operators. They are commutative, associative, increasing and with $\pi_i(x) = 0$ as neutral element. They satisfy the following property:

$$\pi_\vee(x) \geq \max_{i \in [1,m]} \pi_i(x) \tag{12.15}$$

- *Compromise combination*: this corresponds to intermediate *cautious* behaviors. $T(a, b)$ is a compromise combination if it satisfies:

$$\min(a, b) < T(a, b) < \max(a, b) \tag{12.16}$$

On illustrative purpose, we can consider the following toy problem. To estimate how old a person is, two estimates are available, each one modeled by a fuzzy set. These fuzzy sets are represented in Fig. 12.4.a - note that they are highly conflicting. From these two information sources, we want to classify a person into one of the three following classes: young (under 30), middle age (between 30 and 65) or old (above 65). To illustrate the three possible modes of combination, we aggregate the information with the min operator (T-norm), the max operator (T-conorm) and the three different compromise operators. Results are presented in Fig. 12.4. The decision is taken by selecting the class corresponding to the maximum membership.

Conjunctive combination Fig. 12.4.b presents the result obtained with the min operator, *i.e.*, the less severe conjunctive operator. It is a unimodal fuzzy set. This fuzzy set is subnormalized, but this problem could be solved using $\pi'_{\wedge}(x) = \frac{\pi_{\wedge}(x)}{\sup_x(\pi_{\wedge}(x))}$ but this would not change the shape of the result. In this case, the decision would be *middle age*, which is not compatible with any of the initial sources. In this case, the sources strongly disagree and the conjunctive fusion does not help in the classification. As a conclusion, conjunctive operators are not suited for conflicting situations.

Disjunctive combination Fig. 12.4.c presents the result obtained with the max operator, *i.e.*, the less indulgent disjunctive operator. The resulting membership function is multi-modal and each maximum is of equal amplitude. Again, no satisfactory decision can be made.

Compromise combination Three different such operators are discussed. They are all based on the measure of the conflict between sources, defined as $1 - C$ with:

$$C(\pi_1, \pi_2) = \sup_x \min(\pi_1(x), \pi_2(x)) \tag{12.17}$$

The compromise combination operators have been proposed by D. Dubois and H. Prade in [35]. Bloch has classified these operators as *Contextual Dependant (CD) Operators* [36]. Where context can be, e.g, a conflict between the sources, knowledge about reliability of a source, some spatial information. These operators have been proposed in possibility theory [37] but they can also be used in fuzzy set theory for combining membership functions [36]. Being able to adapt to the context, these operators are more flexible and thus provide interesting results. The first considered operator (12.18):

$$\pi(x) = \begin{cases} \max\left(\frac{\min(\pi_1(x),\pi_2(x))}{C(\pi_1,\pi_2)}, \min(\max(\pi_1(x),\pi_2(x)), 1 - C(\pi_1,\pi_2))\right) & if \quad C(\pi_1,\pi_2) \neq 0 \\ \max(\pi_1(x),\pi_2(x)) & if \quad C(\pi_1,\pi_2) = 0 \end{cases} \tag{12.18}$$

adapts its behavior as a function of the conflict between the sources:

- it is conjunctive if the sources have low conflict
- it is disjunctive if the sources have high conflict
- it behaves in a compromise way in case of partial conflict

Fig 12.4.d presents the obtained result using operator (12.18). Corresponding decision (middle age) is still not satisfactory.
In this case, some information on source reliability must be included, and the most reliable source(s) should be privileged in the fusion process. Different situations can be considered:

- It is possible to assign a numerical degree of reliability to each source.
- A subset of sources is reliable, but we do not know which one(s).
- The relative reliability of the sources are known, but with no quantitative values. However, priorities can be defined between the sources.

The two following adaptive operators are examples of *prioritized fusion operator* [35].

$$\pi(x) = \min\left(\pi_1(x), \max\left(\pi_2(x), 1 - C(\pi_1, \pi_2)\right)\right) \tag{12.19}$$

$$\pi(x) = \max\left(\pi_1(x), \min\left(\pi_2(x), C(\pi_1, \pi_2)\right)\right) \tag{12.20}$$

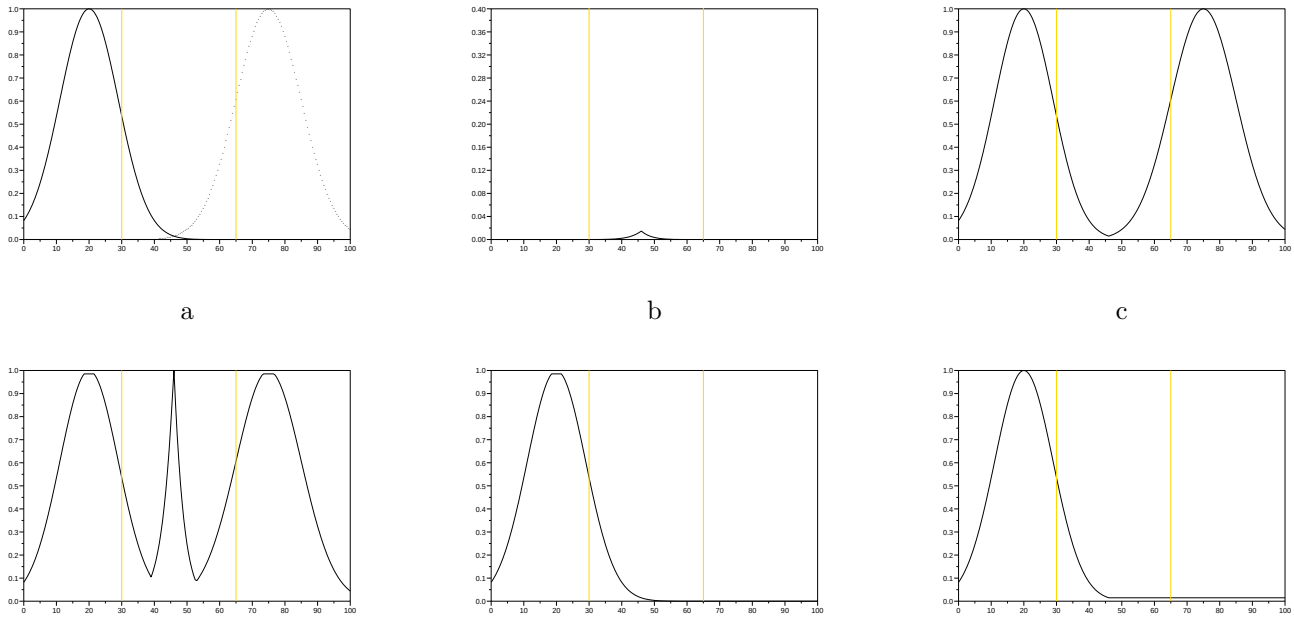


Figure 12.4: Examples of combination operator: a shows the two possibilities distribution, b and c show the result of the min and the max operators, respectively. d,e and f show the results of the three compromise operators presented in (12.18), (12.19) and (12.20), respectively.

For both operators, when $C(\pi_1, \pi_2) = 0$, π_2 contradicts π_1 and only the information provided by π_1 is retained. In this case, π_2 is considered as a specific piece of information while π_1 is viewed as a fuzzy default value. Assuming π_1 is more accurate than π_2, we get the result presented in Fig. 12.4.e and f, enabling a satisfactory decision.

As a conclusion, conjunctive and disjunctive combination operators are ill suited to handle conflicting situations. These situations should be solved by CD operators, incorporating reliability information.

12.3.2 Measure of confidence

Point-wise accuracy

For a given pixel and a given classifier, we propose to interpret the degree of fuzziness of the fuzzy set $\pi_i(x)$ defined in (12.12) as a point-wise measure of the accuracy of the method. We intuitively consider that the classifier is *reliable* if one class has a high membership value while all the others have a membership value close to zero. On the contrary, when no membership value is significantly higher than the others, the classifier is *unreliable* and the results it provides should not be taken too much into account in the final decision. In other words, uncertain results are obtained when the fuzzy set $\pi_i(x)$ has a high fuzziness degree, the highest degree being reached for uniformly distributed membership values.

To reduce the influence of unreliable information and thus enhance the relative weight of reliable information, we weight each fuzzy set by:

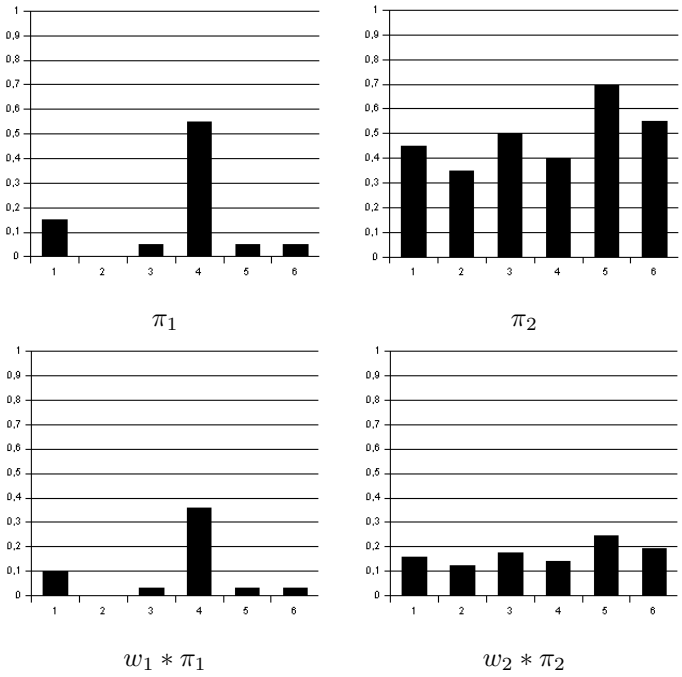


Figure 12.5: Normalization effects. This figure shows two fuzzy sets (π_1 and π_2) with different fuzziness ($H_{\alpha QE}(\pi_1) = 0.51$, $H_{\alpha QE}(\pi_2) = 0.97$, $w_1 = 0.65$ and $w_2 = 0.35$). The normalization effect is shown on the right side. Influence of classifier 2 is more reduced by w_2 than classifier 1 is reduced by w_1.

$$\begin{cases} w_i = \dfrac{\sum_{k=0, k\neq i}^{m} H_{\alpha QE}(\pi_k)}{(m-1)\sum_{k=0}^{m} H_{\alpha QE}(\pi_k)} \\ \sum_{i=0}^{m} w_i = 1 \end{cases} \tag{12.21}$$

where $\alpha = 0.5$, $H_{\alpha QE}(\pi_k)$ is the fuzziness degree of source k, and m is the number of sources. When a source has a low fuzziness degree, w_i is close to 1 and it only slightly affects corresponding fuzzy set. Fig. 12.5 illustrates the effects of this normalization.

Global accuracy

Beyond the adaptation to the local context described in the previous paragraph, we can also use prior knowledge regarding the performances of each classifier. This knowledge is modeled for each classifier i and for each class j by a parameter f_i^j. Such global accuracy can be determined by a separate statistical study on each of the used classifiers. If, for a given class j, the user considers that the results provided by classifier i are satisfactory, parameter f_i^j is set to one. Otherwise, it is set to zero. Since this decision is binary, we assume for each class, that there is at least one method ensuring a satisfactory global reliability.

12.3.3 Combination operator

Numerous combination rules have been proposed in the literature, from simple conjunctive or disjunctive rules, such as min or max operators, to more elaborated CD operators, such as defined by (12.19) and (12.20) where the relative reliability of each source is used. However, with these operators, sources have always the same hierarchy and the

fusion scheme does not adapt to the local context. In [11], we propose the following extension:

$$\mu_f^j(x) = \max\Big(\min\big(w_i\mu_i^j(x), f_i^j(x)\big),\ i \in [1,m]\Big) \tag{12.22}$$

where f_i^j is the global confidence of source i for class j, w_i is the normalization factor defined in (12.21), and μ_i^j is an element of the fuzzy set π_i defined in (12.12). This combination rule ensures that only reliable sources are taken into account for each class (pre-defined coefficients f_i^j), and that the fusion also automatically adapts to the local context by favoring the source that is locally the most reliable (weighting coefficients w_i).

12.4 The Fusion scheme

We present here the complete proposed fusion scheme. In a first step, each classifier is applied separately (but no decision is taken). In a second step, the results provided by the different algorithms are aggregated. The final decision is taken by selecting the class with the largest resulting membership value.

The fusion step is organized as follows:
For each pixel :

1. Separately build the fuzzy set $\pi_i(x) = \{\mu_i^1(x), \mu_i^2(x), \ldots, \mu_i^j(x), \ldots, \mu_i^n(x)\}$ for each classifier i, with n classes.
2. Compute the fuzziness degree $H_{\alpha QE}(\pi_i)$ of each fuzzy set $\pi_i(x)$.
3. Normalize data with w_i defined in (12.21).
4. Apply operator (12.22).
5. Select the class corresponding to the highest resulting membership degree.

The block diagram of the fusion process is given in Fig. 12.6.
Note that in Fig. 12.6, the range of the fuzzy sets is rescaled before the fusion step in order to combine data with the same range. This is achieved with the following range stretching algorithm:

- for all $\pi_i(x) = \{\mu_i^1(x), \ldots, \mu_i^j(x), \ldots, \mu_i^n(x)\}$, compute :
  - $M = \max_{j,x}\left[\mu_i^j(x)\right]$,
  - $m = \min_{j,x}\left[\mu_i^j(x)\right]$,
  - for all $\mu_i^j(x)$, compute :
    - $\mu_i^j(x) = \dfrac{\mu_i^j(x) - m}{M - m}$.

12.5 Experimental Results

12.5.1 Test image

Airborne data from the ROSIS-03 (Reflective Optics System Imaging Spectrometer) optical sensor are used for the experiments. The flight over the city of Pavia, Italy, was operated by the Deutschen Zentrum fur Luft- und Raumfahrt (DLR, the German Aerospace Agency) in the framework of the HySens project, managed and sponsored by the European Union. According to specifications the number of bands of the ROSIS-03 sensor is 115 with a spectral coverage ranging from 0.43 to 0.86μm. The spatial resolution is 1.3m per pixel. The original data set is 610 by 340 pixels. Some channels (12) have been removed due to noise. The remaining 103 spectral dimensions are processed.

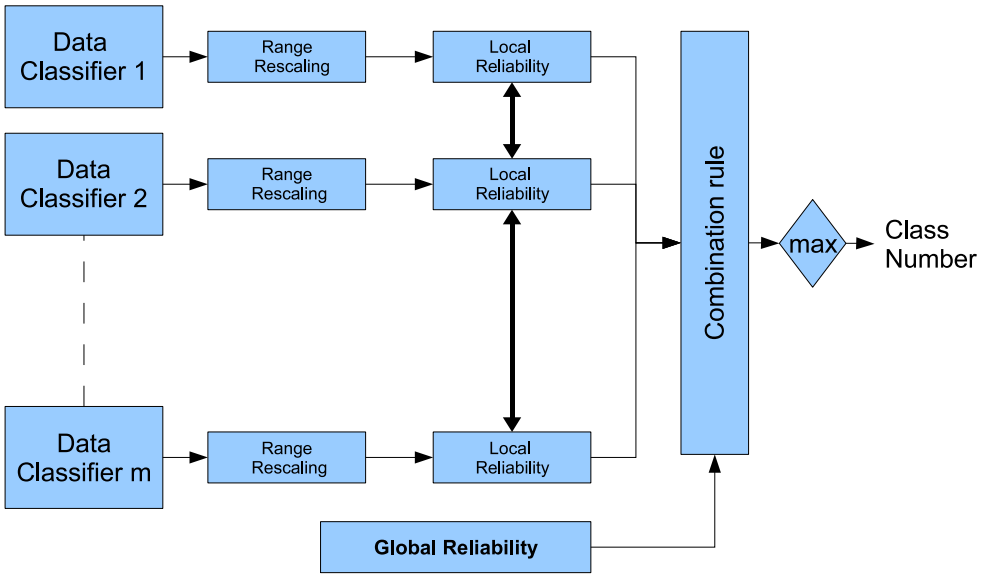
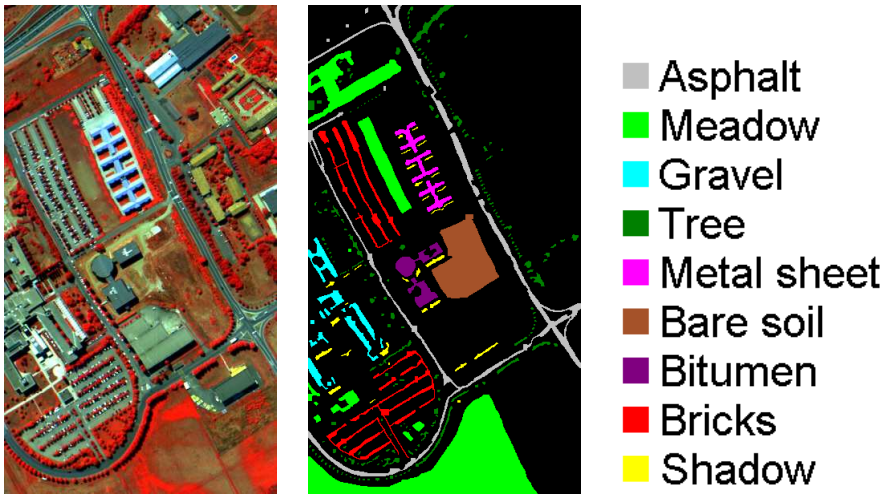


Figure 12.6: Block Diagram of the fusion method

Nine classes of interest are considered, namely: trees, asphalt, bitumen, gravel, metal sheet, shadow, bricks, meadow and soil.

Fig. 12.7(a) presents a three-channel color composite of the original data, where channels 80, 45 and 10 of the original data are used for red, green and blue, respectively. The available reference data is shown in Fig. 12.7(b) and the number of training and test samples is given in Table 12.2.



(a) (b)

Figure 12.7: ROSIS University area: (a) three-channel color composite, channels 80 (red), 45 (green) and 10 (blue), and (b) available reference data and information classes.

12.5.2 Classifier based on mathematical morphology and neural network

In this section, we present a classifier based on a morphological feature extraction, the classification being performed with an artificial neural network. We first briefly recall the concept of granulometry. We then explain how the feature extraction is extended to the case of hyperspectral images.

Table 12.2: Information classes and samples.

Class		Samples	
No.	Name	Train	Test
1	Asphalt	548	6304
2	Meadows	540	18146
3	Gravel	392	1815
4	Tree	524	2912
5	Metal sheet	265	1113
6	Bare Soil	532	4572
7	Bitumen	375	981
8	Brick	514	3364
9	Shadow	231	795
Total		3921	40002

Feature extraction & granulometry

Granulometries are popular and powerful tools derived from the mathematical morphology theory. They are classically used for the analysis of the size distribution of particles in an image. They can be applied in various applications, ranging from the study of porous media to texture segmentation [38]. More information on mathematical morphology can be found in [39] [40]. Ref. [13] is an application oriented book on morphological image analysis, from the principles to recent developments, and Ref. [12] is a survey paper investigating the use of advanced morphological operators in the general frame of satellite remote sensing.

Granulometries have recently been introduced in remote sensing image processing for the classification of urban areas [14] [15]. As a matter of fact, traditional pixel classification techniques used for remotely sensed rural areas turned to be ill-suited for urban areas, especially at high resolution. Beyond the spectral signature, there is actually a strong need for incorporating spatial information in the classification process. That can be achieved using granulometries.

The classical granulometry by opening is obtained by applying morphological opening operations with structuring elements (SE) of increasing size. The consequence is a progressive simplification of the image with a gradual disappearance of the features that are brighter than their immediate neighborhood. Each structure is removed when it becomes smaller than the SE. Using connected operators, such as geodesic reconstruction, no shape noise is introduced: at one given step, every structure is either totally removed or exactly preserved [41]. The opening being an anti-extensive operation, the evolution of the grey level for each pixel can be plotted as a monotonously decreasing curve, from its initial value to a lower bound corresponding to the smallest grey level value in the initial image as the size of the SE increases. Noting $I(x, y)$ as the original image and γ_s the morphological opening by reconstruction using a disc SE of radius s, the morphological profile $MP_\gamma(i)$ is defined for each pixel by:

$$\begin{aligned} MP_\gamma(0) &= I(x,y) \\ MP_\gamma(1) &= \gamma_s[I(x,y)] \\ MP_\gamma(2) &= \gamma_{2s}[I(x,y)] \\ .... &\downarrow .... \\ MP_\gamma(p) &= \gamma_{ps}[I(x,y)] \end{aligned} \tag{12.23}$$

When the size of the SE reaches the characteristic size of one given structure, *i.e.*, when the SE does not fit inside the structure any longer, the structure is removed. Corresponding pixels are then assigned the grey level value of the darker region surrounding the structure. Opening operations only affect structures that are brighter than their

immediate neighborhood. Other structures are left unchanged and thus lead to a constant MP_γ.

Similarly, a granulometry by closing is obtained using morphological closing operations, with the same set of structuring elements. Noting Φ_s the morphological closing by reconstruction with a structuring element of size s, it is given by:

$$MP_\Phi(i) = \Phi is[I(x,y)] \ , \quad i = 1, ..., p \tag{12.24}$$

The closing operation being dual to the opening operation, the closing-based profile provides information regarding the structures that are darker than their immediate surrounding and does not affect structures that are brighter than their surroundings.

Finally, in order to process simultaneously the bright and the dark structures of the image, the two MPs are concatenated:

$$\begin{aligned} MP(i) &= MP_\gamma(i) \qquad for\ i = 1, ..., p \\ &= I(x) \qquad\quad for\ i = 0 \\ &= MP_\Phi(-i) \quad for\ i = -1, ..., -p \end{aligned} \tag{12.25}$$

For illustration, Fig. 12.8 presents one component of the original image (part of Fig. 12.9) and one granulometry with two openings and two closings.

Figure 12.8: Simple morphological profile with 2 opening and 2 closings. In the shown profile, circular structuring elements are used with radius increment 4 ($r = 4,\ 8\ pixels$).

The morphological profile (or its derivate) extracted for every pixel can be used as the input for an artificial network performing the classification [14], potentially by including decision boundary feature extraction (DBFE) to reduce the redundancy in the morphological profile [15].

Extension to hyperspectral data

Similar feature extractions have been proposed to deal with hyperspectral images [16] [17]. One solution consists in first decomposing the data into Principal Components (PC). Fig. 12.9 presents the four first components in the decomposition of the ROSIS original data used in this study. It results in the concentration of the information over a few uncorrelated components. The proposed algorithm is then

- Perform a Principal Component Analysis of the original hyperspectral data
- Select the first Principal Components cumulating 90% of the whole information (the sum of the corresponding eigen values represents 90% of the sum of all the eigen values). In our experiment, this is achieved by selecting the 3 first PCs.

Table 12.3: Confusion Matrix for the Neural Network based Classifier

	asphalt Ref.	meadow Ref.	gravel Ref.	tree Ref.	metal sheet Ref.	bare soil Ref.	bitumen Ref.	brick Ref.	shadow Ref.		
asphalt	6157	1	86	1	1	52	58	36	13		
meadow	44	18476	17	22	1	154	0	49	0		
gravel	0	19	31	4	0	0	0	9	0		
tree	15	23	48	2968	4	9	1	70	8		
metal sheet	0	3	993	6	1334	0	0	1200	1		
bare soil	406	127	0	61	0	4791	17	1	1		
bitumen	8	0	23	0	0	22	1254	0	0		
brick	1	0	901	1	4	1	0	2317	0		
shadow	0	0	0	1	1	0	0	0	924	Average Acc	Overall Acc
%	92.85	99.07	1.48	96.87	99.18	95.27	94.29	62.93	97.57	82.17	89.42

- Compute the morphological profile for each pixel, on each component separately. In our experiment, 10 openings and 10 closings are computed, with circular structuring elements with a 2 pixels increment. For one given component and for every pixel, the extracted vector has a dimension equal to $2 \times 10 + 1 = 21$
- For every pixel, concatenate the profiles obtained with each component. This leads to a 3 components $\times$ 21 values = 63 dimensional feature vector. This is illustrated by Fig. 12.10 in a reduced case.
- This vector is used as the input for the neural network for classification.

Results

Fig. 12.12(a) presents the thematic classification map obtained with this Neural Network based classifier. The corresponding confusion matrix is given by Table 12.3. The classification accuracy is fairly good. However, one class, namely "gravel", has poor results with the corresponding pixels being spread aver various classes and with a confusion with the "metal sheet" class.

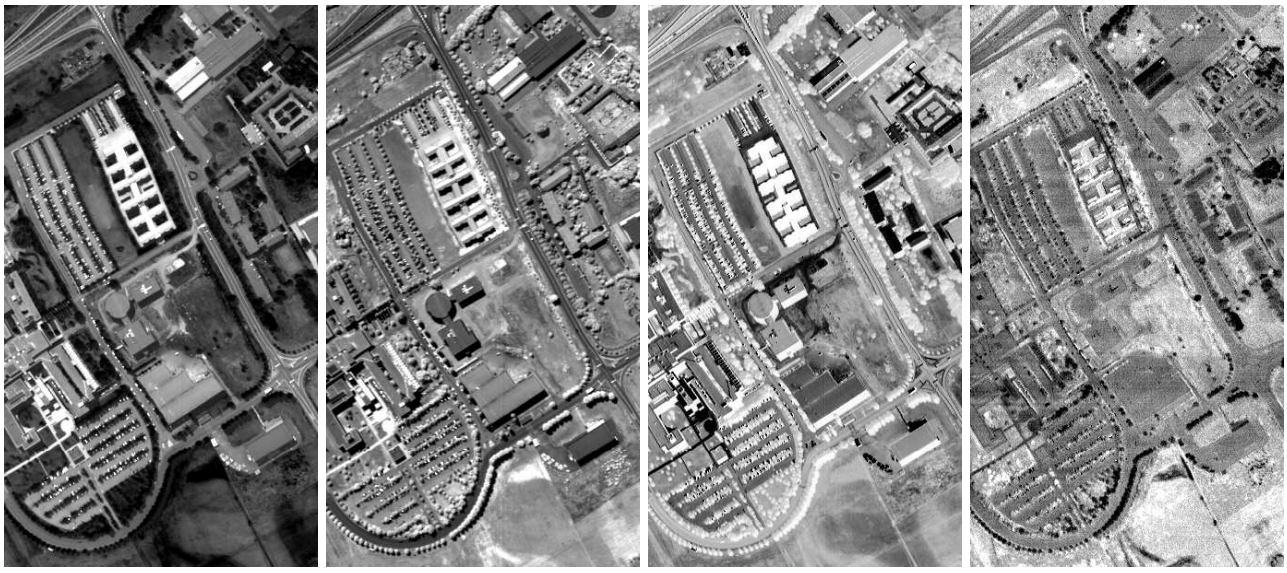


Figure 12.9: ROSIS University area, most important principal components, 1st (left) through 4th (right).

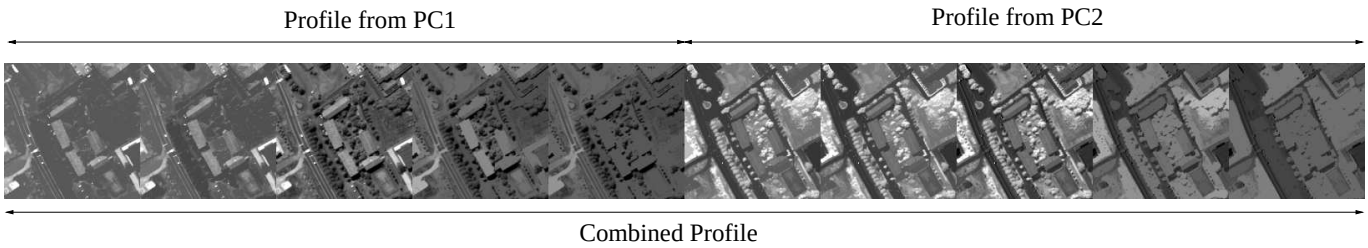


Figure 12.10: Extended morphological profile of two images. Each of the original profile has 2 opening and 2 closings. Circular structuring element with radius increment 4 was used ($r = 4, 8$).

Assessing the point-wise accuracy

For every pixel, the output of the neural network gives the posterior probability corresponding to each class. Consequently, the fuzzy set $\pi_i(x)$ is directly derived and is used as the input to the previously described fusion scheme for this neural network based classifier.

12.5.3 Classifier based on Support Vector Machines

In this section, we present a second classifier which is detailed and discussed in [42]. It is based on Support Vector Machines who are known to be well suited for high dimensional classification problems [3,4]. As stated in the introduction, SVMs characterize classes using a geometrical criterion rather than statistical criteria. They seek a separating hyperplane maximizing the distance with the closest training samples for two classes. This approach allows SVMs to have a very high capability of generalization and, as a consequence, only require a few training samples. For non linearly separable data, SVMs use the kernel trick to map the data onto a higher dimensional space where they are linearly separable [5].

Here, we consider multiclass SVMs without any feature reduction of the original hyperspectral data. The standard Gaussian radial basis kernel with L2-norm distance is used. This algorithm was proved to provide interesting classification accuracy, even in the case of limited training set [42]. Note that other kernels could be considered, such as the spectral angle mapper which basically computes the angle between two vectors in the vector space [10, 42, 43].
In the following, we present the used classifier and start by briefly recalling the general mathematical formulation of SVMs. Starting from the linearly separable case, optimal hyperplanes are introduced, then the classification problem is modified to handle non-linearly separable data and a brief description of multiclass strategies is given. Finally, kernel methods are presented.

Linear SVMs

For a two classes problem in a n-dimensional space $\mathbb{R}^n$, we assume that N training samples, $\mathbf{x}_i \in \mathbb{R}^n$, are available with their corresponding label $y_i = \pm 1$: $\{(\mathbf{x}_i, y_i) \mid i \in [1, N]\}$. The SVMs method consists in finding the hyperplane that maximizes the margin (see Fig. 12.11), *i.e.*, the distance to the closest training data points in both classes. Noting $\mathbf{w} \in \mathbb{R}^n$ as the vector normal to the hyperplane and $b \in \mathbb{R}$ as the bias, the hyperplane H_p is defined as

$$\mathbf{w} \cdot \mathbf{x} + b = 0, \ \forall \mathbf{x} \in H_p \tag{12.26}$$

where $\mathbf{w} \cdot \mathbf{x}$ is the dot product between $\mathbf{w}$ and $\mathbf{x}$. If $\mathbf{x} \notin H_p$ then $f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b$ is the distance of $\mathbf{x}$ to H_p. According to the previous statement, such a hyperplane has to satisfy:

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) > 1, \ \forall i \in [1, N] \tag{12.27}$$

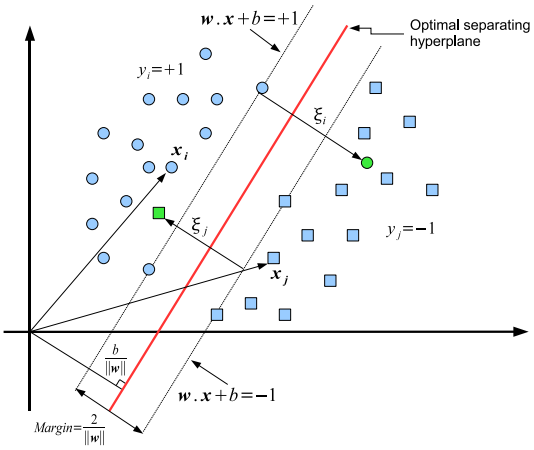


Figure 12.11: Classification of a non-linearly separable case by SVMs. There is one non separable feature vector in each class.

Finally, the optimal hyperplane has to maximize the margin: $2/\|\mathbf{w}\|$, this is equivalent to minimize $\|\mathbf{w}\|/2$ and leads to the following quadratic optimization problem:

$$\min \left[\frac{\|\mathbf{w}\|^2}{2} \right], \text{ subject to (12.27)} \tag{12.28}$$

For non-linearly separable data, *slack* variables ξ are introduced to deal with misclassified samples (see Fig. 12.11). Eq. (12.27) becomes

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) > 1 - \xi_i, \ \xi_i \geq 0 \quad \forall i \in [1, N] \tag{12.29}$$

The final optimization problem turns to:

$$\min \left[\frac{\|\mathbf{w}\|^2}{2} + C \sum_{i=1}^{N} \xi_i \right], \text{ subject to (12.29)} \tag{12.30}$$

where the constant C controls the amount of penalty. It can be solved by considering the dual optimization problem using Lagrange multipliers α:

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) \\ \text{subject to} \quad & 0 \leq \alpha_i \leq C \quad \forall i \in [1, N] \\ & \sum_{i=1}^{N} \alpha_i y_i = 0. \end{aligned} \tag{12.31}$$

Finally:

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i. \tag{12.32}$$

The solution vector is a linear combination of some samples of the training set, whose α_i is non-zero, called *Support Vectors*. The hyperplane decision function can thus be written as:

$$y_u = sgn \left(\sum_{i=1}^{N} y_i \alpha_i (\mathbf{x}_u \cdot \mathbf{x}_i) + b \right) \tag{12.33}$$

where $\mathbf{x}_u$ is an unseen sample.

Multiclass SVMs

SVMs are designed to solve binary problems where the class labels can only take two values: ± 1. For a remote sensing application, several classes are usually of interest. Various approaches have been proposed to address this problem. They usually combine a set of binary classifiers. Two main approaches were originally proposed for a m-classes problem [5].

- **One Versus the Rest:** m binary classifiers are applied on each class against the others. Each sample is assigned to the class with the *maximum* output.

- **Pairwise Classification**: $\frac{m(m-1)}{2}$ binary classifiers are applied on each pair of classes. Each sample is assigned to the class getting the highest number of votes. A vote for a given class is defined as a classifier assigning the pattern to that class.

The *pairwise classification* has shown to be more suitable for large problems [44]. Even though the number of classifiers to handle is larger than for the *one versus the rest* approach, the whole classification problem is decomposed into much simpler ones. This second approach is therefore used in this study.

Nonlinear SVMs

Kernel methods are a generalization of SVMs providing nonlinear decision functions and thus improving classification abilities. Input data are mapped onto a higher dimensional space $\mathbb{H}$ using a nonlinear function Φ:

$$\begin{array}{rcl} \mathbb{R}^n & \rightarrow & \mathbb{H} \\ \mathbf{x} & \rightarrow & \Phi(\mathbf{x}) \\ \mathbf{x}_i \cdot \mathbf{x}_j & \rightarrow & \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j) \end{array} \tag{12.34}$$

The expensive computation of $\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$ in $\mathbb{H}$ is reduced using the *kernel trick* [5]:

$$\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j) = K(\mathbf{x}_i, \mathbf{x}_j) \tag{12.35}$$

The kernel K should fulfill Mercer's condition [4]. Using kernels, we never explicitly work in $\mathbb{H}$, and all the computations are done in the original space $\mathbb{R}^n$.
For classification of remote sensing images, two kernels are popular: the inhomogeneous polynomial function and the Gaussian radial basis function (RBF).

$$K_{POLY}(\mathbf{x}_i, \mathbf{x}_j) = [(\mathbf{x}_i \cdot \mathbf{x}_j) + 1]^p \tag{12.36}$$

$$K_{RBF}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left[-\gamma \left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2\right] \tag{12.37}$$

Radial basis functions can be written as follows [5]: $K(\mathbf{x}_i, \mathbf{x}_j) = f(d(\mathbf{x}_i, \mathbf{x}_j))$ where d is a metric on $\mathbb{R}^n$ and f is a function on $\mathbb{R}_0^+$. For the Gaussian RBF, $f(t) = \exp(-\gamma t^2)$, $t \in \mathbb{R}_0^+$, and $d(\mathbf{x}_i, \mathbf{x}_j)) = \|\mathbf{x}_i - \mathbf{x}_j\|$, *i.e*, the Euclidean distance. As mentioned in [43], Euclidean distance is not scale invariant, however due to atmospheric attenuation or variation in illumination, spectral energy can be different for two samples even if they belong to the same class. To handle such a problematic case, scale invariant metrics can be considered. *Spectral Angle Mapper* (SAM) is a well known scale invariant metric, it has been widely used in many remote sensing problems and it has been shown to be robust to variations in spectral energy [43]. This metric α focuses on the angle between two vectors:

$$\alpha(\mathbf{x}_i, \mathbf{x}_j) = \arccos\left(\frac{\mathbf{x}_i \cdot \mathbf{x}_j}{\|\mathbf{x}_i\| \, . \, \|\mathbf{x}_j\|}\right) \tag{12.38}$$

Table 12.4: Indices of confidence

	Neural Network	SVM
asphalt	1	1
meadow	1	1
gravel	0	1
tree	1	1
metal sheet	1	1
bare soil	1	1
bitumen	1	1
brick	0	1
shadow	1	1

Table 12.5: Confusion Matrix for the SVM based Classifier

	asphalt	meadow	gravel	tree	metal sheet	bare soil	bitumen	brick	shadow		
	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.		
asphalt	5551	0	29	0	0	21	99	35	22		
meadow	32	13100	11	28	0	193	0	14	0		
gravel	113	0	1476	2	0	0	1	186	7		
tree	21	2037	0	2997	2	55	0	6	0		
metal sheet	20	0	0	5	1337	97	0	0	0		
bare soil	16	3493	6	31	0	4639	0	23	0		
bitumen	346	0	5	0	0	0	1218	0	0		
brick	515	19	572	1	0	24	12	3409	3		
shadow	17	0	0	0	6	0	0	0	915	Average Acc	Overall Acc
%	83.71	70.26	70.32	97.81	99.41	92.25	91.58	92.59	96.62	88.28	80.98

See [42] for a comparison of RBF kernels with the Euclidean distance (12.37) and the spectral angle mapper (12.39).

$$K_{SAM}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left[-\gamma\alpha(\mathbf{x}_i, \mathbf{x}_j)^2\right] \tag{12.39}$$

Both kernels fulfill Mercer's conditions and optimal hyperplanes can therefore be found. In this study, we will only consider RBF kernels with the Euclidean distance which turned out to be better suited for the analysis of urban areas.

Results

Fig. 12.12(b) presents the thematic classification map obtained with this Gaussian kernel SVM classifier. The corresponding confusion matrix is given by Table 12.5. Generally speaking, the results obtained are comparable with the results provided by the neural network. The overall accuracy is lower but the average accuracy is higher since the performances are more uniform among the different classes. In particular, this method performs much better than the neural network for the "gravel" class and significantly better for the "bitumen" class as well.

As a conclusion, the two presented classifiers provide complementary information and a fusion of their results should improve the performances. As a prior knowledge for the fusion algorithm, we can define indices of confidence (global accuracy) for the two algorithms. They are presented in Table 12.4): both algorithms are well suited for all the classes, except the neural network for the "gravel" and the "brick" classes.

Assessing the point-wise accuracy

Since a *pairwise classification* has been used, $\frac{m(m-1)}{2}$ binary classifiers are evaluated, thus it is not possible to derive directly a fuzzy set $\pi_i(x)$ as with the neural network based algorithm. Probabilities are built following a standard

method [45]: for each considered class, the obtained number of votes is divided by the total number of votes, *i.e.* the number of pairwise classifiers (*i.e.* $\frac{m(m-1)}{2}$). For every pixel, it results in the construction of a fuzzy set where a probability value ranging from 0 to 1 is associated to each class. These probabilities are used as the input to the previously described fusion scheme for the SVM based classifier.

(a) (b)

Figure 12.12: Result of the classification using the Neural Network (a) and the Support Vector Machine (b).

12.5.4 Decision fusion

In this section, we present the results obtained with the different fusion operators presented in section 12.3.

Fig. 12.13(d) and (e) present the thematic maps obtained with the *min* and the *max* operator, respectively. Corresponding confusion matrices are given in Tables 12.6 and 12.7, respectively. This simple operators do not lead to any improvement, the class "tree" being even completely lost and confused with "asphalt". This underlines the need for a proper handling of conflictual situations.

Fig. 12.13(a), (b) and (c) present the thematic maps obtained with the operators from Eq.(12.18), (12.19) and (12.20), respectively. Corresponding confusion matrices are given in Tables 12.8, 12.9 and 12.10, respectively. For these operators, the measure of conflict between the different classifiers is given by Eq. 12.17. For the operators (12.19) and (12.20), the neural network classifier is chosen as the priority in case of conflict since this is the one providing the best overall accuracy. It is striking to note that operators (12.18) and (12.19) actually lead to the same results as operator *min* (they are almost identical). This is due to the measure of conflict $1 - C$ that is most of the times quite low: with values of C higher than 0.6, operators (12.18) and (12.19) indeed converge towards the *min* operator.
None of these operators provides fully satisfactory results. As a matter of fact, more flexibility is needed in the fusion of conflictual situations. This is achieved with the proposed adaptive fusion scheme whose results are presented on Fig. 12.13(f) and in Table 12.11. Though providing quantitative results close to the results obtained with the SVM, the thematic map is less noisy, and no class has bad results, unlike with the neural network ("gravel") or the other fusion operators ("tree").

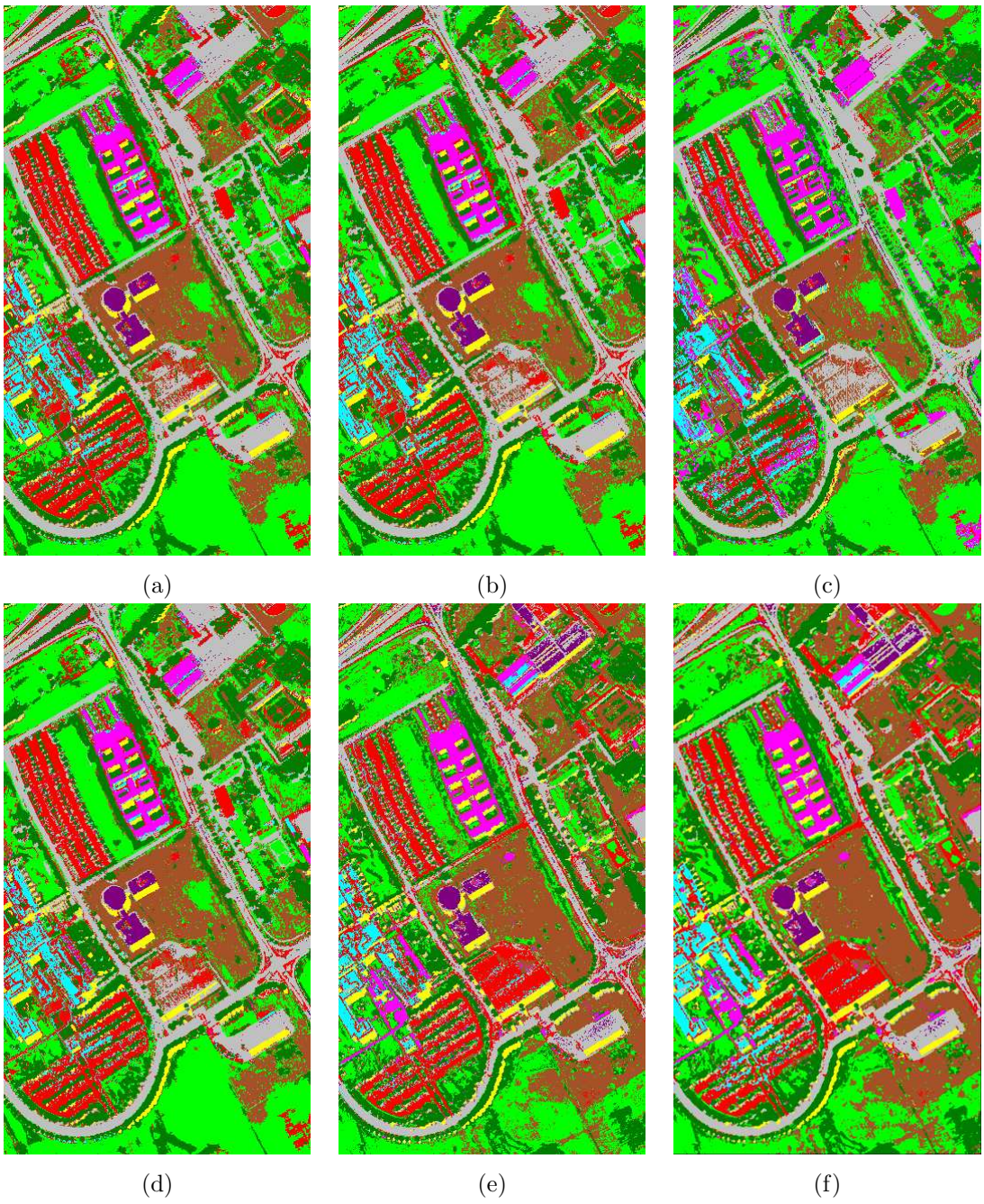


Figure 12.13: Result of the classification with a decision fusion using various operators. (a) operator (12.18), (b) operator (12.19), (c) operator (12.20), (d) operator *min*, (e) operator *max* and (f) proposed adaptive operator.

12.6 Conclusion

In this chapter, we have presented two classifiers dealing with hyperspectral images. The first one uses operators derived from mathematical morphology to extract relevant features. The classification is performed using an artificial neural network. The second one classifies each pixel directly from its spectral value using a kernel Support Vector Machine.

To take advantage of these two classifiers and their complementary properties, a general decision fusion framework, based on a fuzzy combination rule, is presented. The proposed method is adaptive and enables a satisfactory handling of the conflictual situations where the different classifiers disagree. Two measures of accuracy are used in the combination rule: the first one, based on prior knowledge, defines global reliabilities, both for each classifier and each class. The

Table 12.6: Confusion Matrix for the decision fusion using the *min* operator

	asphalt Ref.	meadow Ref.	gravel Ref.	tree Ref.	metal sheet Ref.	bare soil Ref.	bitumen Ref.	brick Ref.	shadow Ref.		
asphalt	6114	1857	2	2948	23	53	82	26	94		
meadow	26	16496	0	59	0	902	0	4	0		
gravel	25	0	1521	3	59	5	1	256	70		
tree	0	0	0	0	0	0	0	0	0		
metal sheet	0	0	0	0	1198	7	0	0	0		
bare soil	28	189	0	47	0	3959	0	0	0		
bitumen	158	0	0	0	0	0	1245	0	0		
brick	278	107	576	6	48	103	2	3396	31		
shadow	2	0	0	1	17	0	0	0	752	Average Acc	Overall Acc
%	92.20	88.46	72.46	0	89.07	78.72	93.61	92.23	79.41	76.24	81.08

Table 12.7: Confusion Matrix for the decision fusion using the *max* operator

	asphalt Ref.	meadow Ref.	gravel Ref.	tree Ref.	metal sheet Ref.	bare soil Ref.	bitumen Ref.	brick Ref.	shadow Ref.		
asphalt	5572	1921	13	3011	2	52	97	38	18		
meadow	46	13932	9	29	0	193	0	12	0		
gravel	110	0	1485	0	0	0	1	186	8		
tree	0	0	0	0	0	0	0	0	0		
metal sheet	17	2	5	5	1337	98	1	0	1		
bare soil	19	2775	6	19	0	4665	1	32	2		
bitumen	352	0	3	0	0	0	1219	9	0		
brick	500	19	578	0	0	21	11	3405	1		
shadow	15	0	0	0	6	0	0	0	917	Average Acc	Overall Acc
%	84.03	74.71	70.75	0	99.41	92.76	91.65	92.48	96.83	78.07	76.05

second one automatically estimates the point wise reliability of the results provided by each classifier and, thus, enables the adaptation of the fusion rule to local context. The proposed approach does not need any training and the computational load remains low. As a result, a better classification is obtained, providing less noisy thematic maps.

On should underline that no prior assumption is needed regarding the modeling of the data (e.g, Bayes theory, possibility theory, *etc*) before the data are fused. A key point lies in the generality of the presented framework for *decision level fusion*. Though only two classifiers were used in this study, additional algorithms could easily be added to the process. For instance, dedicated algorithms such as street trackers could be used without increasing errors in the other classes.

In this chapter, the α-Quadratic entropy was chosen for the fuzziness evaluation because the sensibility of that measure can be modified with the value of α. Note that several other measures can be used, e.g., the *fuzzy entropy* [29].

One limitation of the proposed approach is the use of binary values for the global confidence. With fuzzy confidence, the combination rule could be rewritten with *T-conorm* and *T-norm*, both of which are less indulgent and less severe than *max* and *min*. Moreover, the use of the *T-conorm* and *T-norm* would allow a finer definition of global accuracy.

Acknowledgment

This research was supported in part by the Research Fund of the University of Iceland and the Jules Verne Program of the French and Icelandic governments (PAI EGIDE).

Table 12.8: Confusion Matrix for the decision fusion using operator (12.18)

	asphalt Ref.	meadow Ref.	gravel Ref.	tree Ref.	metal sheet Ref.	bare soil Ref.	bitumen Ref.	brick Ref.	shadow Ref.		
asphalt	6114	1857	2	2948	23	53	82	26	94		
meadow	26	16496	0	59	0	902	0	4	0		
gravel	25	0	1521	3	59	5	1	256	70		
tree	0	0	0	0	0	0	0	0	0		
metal sheet	0	0	0	0	1198	7	0	0	0		
bare soil	28	189	0	47	0	3959	0	0	0		
bitumen	158	0	0	0	0	0	1245	0	0		
brick	278	107	576	6	48	103	2	3396	31		
shadow	2	0	0	1	17	0	0	0	752	Average Acc	Overall Acc
%	92.20	88.46	72.46	0	89.07	78.72	93.61	92.23	79.41	76.24	81.08

Table 12.9: Confusion Matrix for the decision fusion using operator (12.19)

	asphalt Ref.	meadow Ref.	gravel Ref.	tree Ref.	metal sheet Ref.	bare soil Ref.	bitumen Ref.	brick Ref.	shadow Ref.		
asphalt	6116	1857	2	2948	23	63	82	26	94		
meadow	26	16496	0	59	0	898	0	4	0		
gravel	23	0	1521	3	59	5	1	256	70		
tree	0	0	0	0	0	0	0	0	0		
metal sheet	0	0	0	0	1198	7	0	0	0		
bare soil	28	189	0	47	0	3959	0	0	0		
bitumen	158	0	0	0	0	0	1245	0	0		
brick	278	107	576	6	48	97	2	3396	31		
shadow	2	0	0	1	17	0	0	0	752	Average Acc	Overall Acc
%	92.23	88.46	72.46	0	89.07	78.72	93.61	92.23	79.41	76.24	81.08

Table 12.10: Confusion Matrix for the decision fusion using operator (12.20)

	asphalt Ref.	meadow Ref.	gravel Ref.	tree Ref.	metal sheet Ref.	bare soil Ref.	bitumen Ref.	brick Ref.	shadow Ref.		
asphalt	6023	1688	96	2884	10	72	60	367	107		
meadow	216	16164	6	67	123	1170	3	19	8		
gravel	38	17	1347	15	9	1	2	1028	121		
tree	0	0	0	0	0	0	0	0	0		
metal sheet	7	194	379	16	1130	19	4	55	4		
bare soil	83	490	10	58	18	3676	16	38	137		
bitumen	167	0	1	2	9	2	1244	0	0		
brick	94	96	260	12	10	82	1	2175	80		
shadow	3	0	0	10	36	7	0	0	490	Average Acc	Overall Acc
%	90.83	86.67	64.17	0	84.01	73.10	93.53	59.07	51.74	67.01	75.39

Table 12.11: Confusion Matrix for the decision fusion using the proposed adaptive operator

	asphalt Ref.	meadow Ref.	gravel Ref.	tree Ref.	metal sheet Ref.	bare soil Ref.	bitumen Ref.	brick Ref.	shadow Ref.		
asphalt	6370	3681	356	0	36	229	98	232	80		
meadow	2	12272	2	16	0	26	0	8	0		
gravel	0	0	1350	0	0	0	0	27	1		
tree	0	159	0	3041	0	0	0	0	0		
metal sheet	2	0	0	4	1306	78	0	0	0		
bare soil	1	2535	1	1	0	4692	0	9	0		
bitumen	35	0	0	0	0	0	1232	3	0		
brick	221	2	390	0	0	4	0	3403	0		
shadow	0	0	0	2	3	0	0	0	866	Average Acc	Overall Acc
%	96.06	65.81	64.32	99.25	97.10	93.30	92.63	92.42	91.45	88.04	80.73

Bibliography

[1] C. Lee and D. A. Landgrebe, "Analysing high dimensional multispectral data," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 31, pp. 792–800, July 1993.

[2] G. F. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. on Information Theory*, vol. IT-14, pp. 55–63, January 1968.

[3] V. Vapnik, *Statistical Learning Theory*. Wiley, New York, 1998.

[4] C. Burges, "A tutorial on support vector machines for pattern recognition," *Data mining and knowledge discovery*, vol. 2, pp. 121–167, 1998.

[5] B. Scholkopf and A. J. Smola, *Learning with Kernels*. MIT Press, 2002.

[6] J. A. Gualtieri and S. Chettri, "Support vector machines for classification of hyperspectral data," in *Geoscience and Remote Sensing Symposium*, vol. 2. IGARSS '00. Proceedings, July 2000.

[7] G. H. Halldorsson, J. A. Benediktsson, and J. R. Sveinsson, "Support vector machines in multisource classification," in *Geoscience and Remote Sensing Symposium*, vol. 3. IGARSS '03. Proceedings, July 2003.

[8] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 42, pp. 1778–1790, August 2004.

[9] G. F. Foody and A. Mathur, "A relative evaluation of multiclass image classification by support vector machines," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 42, pp. 1335–1343, June 2004.

[10] M. L. G. Mercier, "Support vector machines for hyperspectral image classification with spectral-based kernels," in *Geoscience and Remote Sensing Symposium*, vol. 1. IGARSS '03. Proceedings, July 2003.

[11] M. Fauvel, J. Chanussot, and J. A. Benediktsson, "Decision fusion for the classification of urban remote sensing images," *submitted to IEEE Trans. on Geoscience and Remote Sensing (in revision)*, 2006.

[12] P. Soille and M. Pesaresi, "Advances in mathematical morphology applied to geoscience and remote sensing," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 40, no. 9, pp. 2042–2055, 2002.

[13] P. Soille, *Morphological image analysis, principles and applications, 2nd edition*. Springer, 2003.

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

[15] J. A. Benediktsson, M. Pesaresi, and K. Arnason, "Classification and feature extraction for remote sensing images from urban areas based on morphological transformations," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 41, no. 9, pp. 1940–1949, 2003.

[16] F. Dell'Acqua, P. Gamba, A. Ferrari, J. Palmason, J. Benediktsson, and K. Arnason, "Exploiting spectral and spatial information in hyperspectral urban data with high resolution," *IEEE Geoscience and Remote Sensing Letters*, vol. 1, no. 4, pp. 322–326, 2004.

[17] J. A. Benediktsson, J. A. Palmason, and J. R. Sveinsson, "Classification of hyperspectral data from urban areas based on extended morphological profiles," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 43, no. 3, pp. 480–491, 2005.

[18] A. Plaza, P. Martinez, J. Plaza, and R. Perez, "Dimensionality reduction and classification of hyperspectral image data using sequences of extended morphological transformations," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 43, no. 3, pp. 466–479, 2005.

[19] J. A. Benediktsson and I. Kanellopoulos, "Classification of multisource and hyperspectral data based on decision fusion," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 37, pp. 1367–1377, May 1999.

[20] B. Jeon and D. A. Landgrebe, "Decision fusion approach for multitemporal classification," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 37, pp. 1227–1233, May 1999.

[21] G. Lisini, F. Dell'Acqua, G. Triani, and P. Gamba, "Comparison and combination of multiband classifiers for landsat urban land cover mapping," in *Geoscience and Remote Sensing Symposium*, vol. CD ROM. IGARSS '05. Proceedings, August 2005.

[22] F. Tupin, I. BLoch, and H. Maitre, "A first step toward automatic interpretation of SAR images using evidential fusion of several structure detectors," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 37, no. 3, pp. 1327–1343, May 1999.

[23] F. Tupin and M. Roux, "Detection of building outlines based on the fusion of SAR and optical features," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 58, pp. 71–82, 2003.

[24] J. Chanussot, G. Mauris, and P. Lambert, "Fuzzy fusion techniques for linear features detection in multitemporal sar images," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 37, no. 3, pp. 1292–1305, may 1999.

[25] R. R. Yager, "A general approach to the fusion of imprecise information," *International Journal Of Intelligent Systems*, vol. 12, pp. 1–29, 1997.

[26] L. A. Zadeh, "Fuzzy sets," *Information and Control*, pp. 338–353, 1965.

[27] G. J. Klir and B. Yuan, *Fuzzy set and Fuzzy Logic: Theory and Applications.* Prentice Hall PTR, 1995.

[28] K. Tanaka, *An introduction to Fuzzy Logic for practical application.* Springer, 1996.

[29] A. D. Luca and S. Termini, "A definition of non-probabilistic entropy in the setting of fuzzy set theory," *Information and Control*, pp. 301–312, 1972.

[30] L. A. Zadeh, "Probability measures of fuzzy events," *J.Math. Analysis and Application*, vol. 23, pp. 421–427, 1968.

[31] B. R. Ebanks, "On measures of fuzziness and their representations," *J.Math. Analysis and Application*, vol. 94, pp. 421–427, 1983.

[32] C. Bezdeck, "Measuring fuzzy uncertainty," *IEEE Trans. on Fuzzy Systems*, pp. 107–118, 1994.

[33] I. Bloch, *Fusion d'informations en traitement du signal et des images*, H. Sciences, Ed. 11 rue Lavoisier, 75008 Paris: Germes LAVOISIER, 2003.

[34] M. Oussalah, "Study of some algebraical properties of adaptive combination rules," *Fuzzy set and systems*, vol. 114, pp. 391–409, 2000.

[35] H. Prade and D. Dubois, "Possibility theory in information fusion," *Proceedings of the Third International Conference on Information Fusion*, 2000.

[36] I. Bloch, "Information combination operators for data fusion: A comparative review with classification," *IEEE Trans. on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 26, no. 1, pp. 52–67, January 1996.

[37] D.Dubois and H.Prade, *Combination of information in the framework of possibility theory*, M. A. A. *et al.*, Ed. New York: Academic, January 1992.

[38] R. Gonzalez and R. Woods, *Digital image processing, 2nd edition*. Prentice Hall, 2002.

[39] J. Serra, *Mathematical morphology, Theoretical advances, vol. 2*. Academic Press, 1988.

[40] E. Dougherty, *Mathematical morphology in image processing*. Marcel Dekker, 1993.

[41] J. Crespo, J. Serra, and R. Schafer, "Theoretical aspects of morphological filters by reconstruction," *Signal Processing*, vol. 47, pp. 201–225, 1995.

[42] M. Fauvel, J. Chanussot, and J. A. Benediktsson, "Kernels Evaluation for multiclass classification of hyperspectral remote sensing data," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Toulouse, France, 2006.

[43] N. Keshava, "Distance metrics and band selection in hyperspectral processing with application to material identification and spectral librairies," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 42, pp. 1552–1565, July 2004.

[44] C. W. Hsu and C. J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. on Neural Networks*, vol. 13, pp. 415–425, March 2002.

[45] T. Wu, C. Lin, and R. Weng, "Probability estimates for multi-class classification by pairwise coupling," *Journal of Machine Learning Research*, vol. 5, pp. 975–1005, 2004.