Feature space optimization prior to fuzzy image classification

François Leduc
Defence R&D Canada –Valcartier
2159 Pie-IX North Blvd.
Val-Bélair, Qc, G3J 1X5
Canada
Francois.Leduc@drdc-rddc.gc.ca

Abstract – *This paper presents a method for features space optimization in a context of fuzzy image classification. Based on membership functions intersections, the method allows to select the most appropriate features for objects discrimination. Comparison of the eCognition nearest neighbor algorithms and fuzzy classification is provided with the use of un-optimized and optimized features sets.*

Keywords: feature space optimization, dimensionality reduction, fuzzy sets, classification.

1 Introduction

It is well known that spectral information alone is not sufficient to discriminate objects of interest with multispectral imagery. This is even more true today with high resolution spectral images such as Ikonos and QuickBird data available on the market. Hence, in order to improve objects separability, discriminant features have been proposed such as first and second-order texture measurements [1],[2], fuzzy contours [3], morphologic indices [4] and transformations like normalized difference vegetation index (NDVI) and Tasseled Cap transform [5].

Considering, for example, an initial 4-bands Ikonos multispectral image, we can compute second-order texture measurement using a 5x5 pixels neighborhood with the entropy parameter. This can be applied to the four initial bands thus creating four new features. The same texture measurement can be applied using a 7x7 window thus leading to four other new features. What about the correlation texture parameter ? And what about first-order statistics ? It is easy to see here that the number of features can dramatically increase, leading to slow computation time without necessarily increasing classification results. Of course, a prior knowledge of texture parameters permits to avoid the use of redundant features but still after a first selection of features, it is difficult to evaluate each feature contribution in a classification results.

This study is concerned with the evaluation of features potential for discriminating objects prior to classification. The structure of this paper is as follow. Section 2 presents the classification paradigm while section 3 presents some feature space reduction methods. Section 4 contains a presentation our method and some elements of the fuzzy sets theory. Section 5 presents our methodology and results are shown in section 6. Section 7 discusses the method and the results and finally section 8 concludes this paper.

2 The classification paradigm

Classification can be defined as the association of a land use/land cover attribute to every pixels of an image [6]. There exists a plethora of classification methods going from traditional statistical approaches to A.I. techniques such as neural networks [7].

Traditionally classification was performed using multispectral data such as SPOT-HRV (3 bands for SPOT-1 ad SPOT-2) and Landsat-TM (6 bands for Landsat-4 and Landsat-5 without the thermal band). With higher spatial resolution sensors such as Ikonos and Quickbird, we gain to use other measurements than just spectral bands. Fig. 1 shows an example of multispectral data (Fig. 1-A) and other features extracted from this initial data set. If a classifier is used with only the three spectral bands of Fig. 1-A it might have difficulties to differentiate between roofs and vegetation as both objects have similar spectral properties. By classifying the images using the 6 features of Fig. 1 (the three spectral bands of Fig. 1-A, the NDVI, the Sobel image and the texture image) it can be easier to discriminate between objects of interest.

Fig. 2 shows the results of an unsupervised k-means clustering. We can see that results differs according to the data set used. Fig. 2-A shows the confusion between roofs and vegetation when classifying only the spectral bands. By adding the NDVI the distinction can be done (Fig. 2-B). We can also see that classification B and C and almost similar. This indicates that texture (dissimilarity) does not provide useful information. Classification B has been performed with four features while classification C used 5 features. It becomes important to evaluate the usefulness of each feature in order not to carry too large data sets for no reason. Moreover, in a context of supervised classification, it is useful to evaluate features pertinence according to objects of interest. The next section will present some methods for feature space reduction.

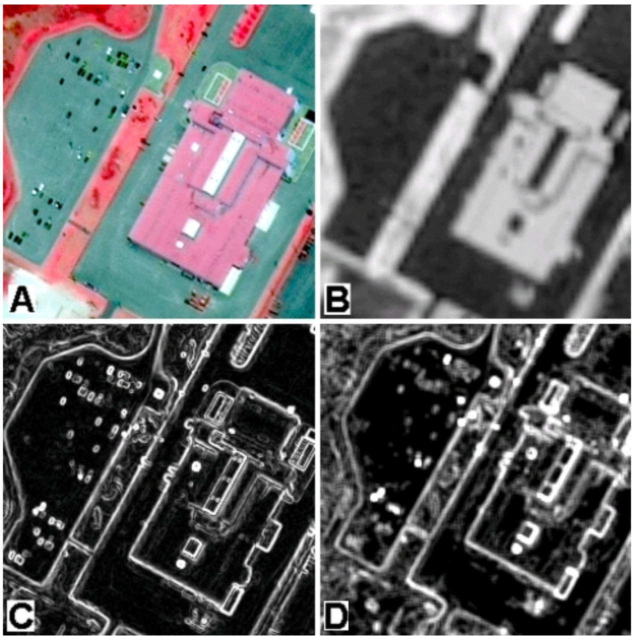

Fig. 1 – A: Multispectral Ikonos image (false-color composite, 1-m Brovey transform). B: NDVI. C: Sobel filtering. D: Dissimilarity (7x7) cooccurence texture measurement. C and D are computed on the NIR band.

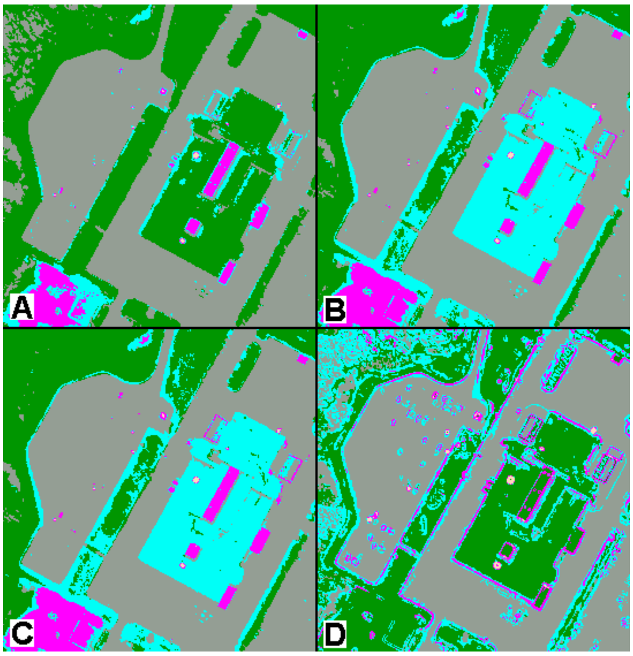

Fig. 2 – Unsupervised classification. A: Using the three spectral bands NIR, red and green. B: Using the three spectral bands and the NDVI. C: using the three spectral bands, the NDVI and texture. D: Using the three spectral bands, the NDVI and the Sobel image.

In this study, we use the term feature to include spectral bands, vegetation indices, texture measurements; in other words any 2-dimensional features that can be represented as images.

3 Feature space reduction

One of the most common used feature space reduction method is the well known Principal components analysis (PCA) [6], [9] also known as Hotelling transform. The aim of this technique is to reduce feature dimensionality by maximizing data variance in a minimum number of new features. Here the quantity of information is related to data variance. When computing a PCA with a whole data set, it might maximize the quantity of information (in term of data variance) in a minimum number of bands but it does not necessarily maximize the information in terms of objects of interest. To avoid this, we might compute a PCA by selecting training sites but still the features space will be optimized by having the maximum data variability in the first component. It happens sometimes that the maximum of information does not necessarily correspond to maximum of information. For example, a clear lake that appears uniformly black in a radar image will not be characterized by large variance. Hence, a PCA could not be used to select the best band for extracting lakes. Finally, another inconvenient with the PCA is that the new features space might be difficult to interpret and with another data set, results might be different. This is a problem of reproducibility.

Pratt [10] divides features space reduction method in two categories: the prototype performance approach and the figure of merit. In the first case, the process consists in classifying data using different data combinations and to keep the data sets that gave the best results. The inconvenient with the prototype performance is that many features combinations are possible and that the measure of performance is also dependant on the classification method. In the second case, the method is the comparison of objects' separability by some criteria such as the Bhattacharyya distance or Mahalanobis distance [11]. Here separability is evaluated by comparing objects by pairs for each feature used. By comparing objects by pairs, one can decide to keep the bands that give the greatest mean divergence. Another decision criterion might be that for each band, the minimum divergence is kept and after that, the bands giving the highest minimums are kept. This can be resumed as keeping the best of the worst separability.

Landgrebe [6] presents other space reduction methods such as Discriminant analysis feature extraction (DAFE). The aim with the DAFE is to maximize the between-class variance and to minimize the average within-class variance. The author mentions that this method is not reliable if classes vector means are very close. He also mentions that it produces optimal features up to the number of classes minus one.

Among other feature space optimization method we can mention the canonical correspondence analysis [12] and the discriminant analysis [13]. Both methods are mainly used to establish a relationship between dependant variables and independent (explanatory) variables.

4 Fuzzy sets-based separability method

Fuzzy sets theory has been proposed by Zadeh in 1965 [14] as a tool for representing uncertainty in knowledge. Fuzzy sets can be used to represent the ambiguity about features characterizing objects. For, example, in the short-wave infrared, the reflectance of conifer is about 11%. But what happens if the measured reflectance is 13% ? Should we reject the belief in being in presence of conifers. We also should keep in mind that we just implicitly used fuzzy concept by saying that conifers' reflectance is *"about"* 11%. This is actually the interest of fuzzy sets because they consider that knowledge about objects is not represented by a single value or by crisp intervals.

Fig. 3 shows examples of fuzzy sets or membership functions according to two features A and B for two objects X and Y. If measuring a feature A value of 26, we get a membership value of 1 to object Y and 0 to object X ($\mu_Y = 1$, $\mu_X = 0$). We then can be sure in the occurrence of object Y. If measuring, at the same time, a feature B value of 18, we obtain the two following membership values: $\mu_Y = 0.2$, $\mu_X = 0.81$. Considering the two features, what will be the final membership for the two objects ?

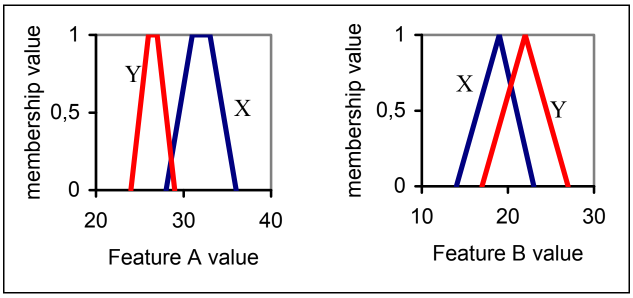

Fig. 3 – Examples of membership functions for two objects X, Y according to two features A, B.

When using information from several sources (sensors, bands, features, experts, etc.) a fusion mechanism is necessary. Fusion with fuzzy sets in mainly done with the use of t-norms and T-conorms [15] such as MIN and MAX operators. The conjunctive fusion (min operator) is well described in [16] where the fusion of *k* sources concerning *m* classes can be expressed as follow:

$$\mu_{Ci}(x) = \min_k[\mu_{Ci,1}(x), \mu_{Ci,2}(x), ..., \mu_{Ci,k}(x)] \qquad (1)$$

where μ_{Ci} is the membership value of pixel x to class C_i. The winning class is then the one that has the highest membership value:

$$\mu_{Cm0}(x) = \max[\mu_{C1}(x), \mu_{C2}(x), ..., \mu_{Cm}(x)] \qquad (2)$$

With the previous example the decision would be taken as follow:

$\mu_Y = \min[1.0 , 0.2] = 0.2$ / $\mu_X = \min[0 , 0.81] = 0$
decision = max[0.2 , 0] = 0.2 (object Y)

The conjunctive fusion can be resumed by first keeping the minimum membership values for each class and then to keep the maximum of these minimums as the final decision.

This min-max approach is a simple fusion method but it considers each feature as having the same reliability for objects discrimination. With the membership functions of Fig. 3 do feature A and B have the same reliability in separating X from Y ? One way to evaluate this could be by measuring the height (h) of fuzzy sets intersections. h can have its values varying between 0 (perfect level of discrimination) to 1 (total confusion). If fuzzy sets intersect at two points, h is given by the maximum values. With feature A (Fig. 3), fuzzy sets intersect at h = 0.2 and with feature B, they intersect at h = 0.66. In that case, it would mean that feature A is better than B to discriminate between X and Y.

This method of evaluating features reliability is simple and is done with consideration of objects of interest. With the example of Fig. 3 the process was trivial but when considering more than twenty features and about 10 classes, the classification process can be long and the results might not be satisfying.

The consideration of h as a measure of separability is inspired by the consensus concept of the possibility theory [19]. The difference is that in the possibility theory, the height of intersection between two possibility distributions represents the level of agreement between two sources. Here the height of intersection represents the degree of similarity (i.e. confusion) between some descriptive features represented by fuzzy sets. The difference can be explained by the fact that in the possibility theory, the information sources use the same universal set while here source have different universal set (reflectance, NDVI, texture, shape, etc.)

The next section presents our methodology and two different ways to compute h.

5 Methodology

In order to evaluate our feature space optimization (FSO) method, we used one Ikonos scene composed of one 1-m panchromatic band and four 4-m multispectral bands. The preprocessing steps, in order, were: 1) bilinear resampling of the four multispectral bands to 1-m spatial resolution; 2) orthorectification of the data set using ERDAS OrthoBase; 3) atmospheric corrections of the 1-m data set using ERDAS ATCOR2; 4) pan-sharpening of the multispectral bands by a Brovey transform [8].

A region of interest of 592 by 501 pixels was extracted (Fig. 4). The four 1-m multispectral bands were segmented with eCognition, which offers a robust bottom-up region-merging segmentation method. This allowed to analyze imagery at the object level instead of the pixel level and to characterize objects with radiometry, shape, size, texture, etc. The segmentation of Fig. 4 produced 1069 polygons.

Fig. 4 – Pan-sharpened Ikonos scene subset. Near infrared, red and green displayed in RBG. 592 x 501 pixels.

Nine classes of interest were defined: 1- helicopter (H), 2- grass (G), 3- forest (F), 4- water (W), 5-tarmac (T), 6- buildings (B), 7- parkings (P), 8- roads (R), 9- soils (S). Fig. 5 shows the ground truth for this subset. This land use map was obtained by classifying the image with almost as many training sites as the number of polygons.

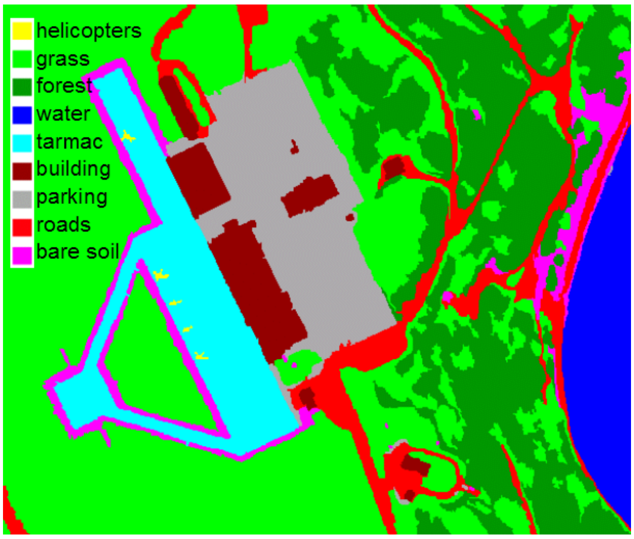

Fig. 5 – Classified image used as ground truth.

Among all features that can be computed with eCognition, 34 features were chosen plus the four multispectral bands and one normalized difference vegetation index (NDVI). The set of 39 features included Haralick texture measurements, polygons size and shape, etc. We don't explain, here, all the features although the list is presented in appendix 1. The interested reader can consult the eCognition documentation for features description. The important point to remember is that objects can be described with many features. Recall that because image is segmented, object can be described by the mean reflectance, their standard deviation, their size, etc.

Training sites (Fig. 6) were identified of the image and allowed to compute statistics describing each class for each feature. Fuzzy sets, or possibility distributions, were built based on histograms. Triangular and trapezoidal shapes were used (Fig. 7). Triangular fuzzy sets were centered on the mean value and delimited each side by two standard deviations. Trapezoidal fuzzy sets were centered on the mean and delimited as follow:

D = mean – 1.5 stdev
E = mean – 0.25 stdev
F = mean + 0.25 stdev
G = mean + 1.5 stdev

These fuzzy sets configurations were really close to the gaussian estimation of the distributions. Also the fuzzy sets shapes were compared to histograms distributions in order to ensure that there was a coherence .

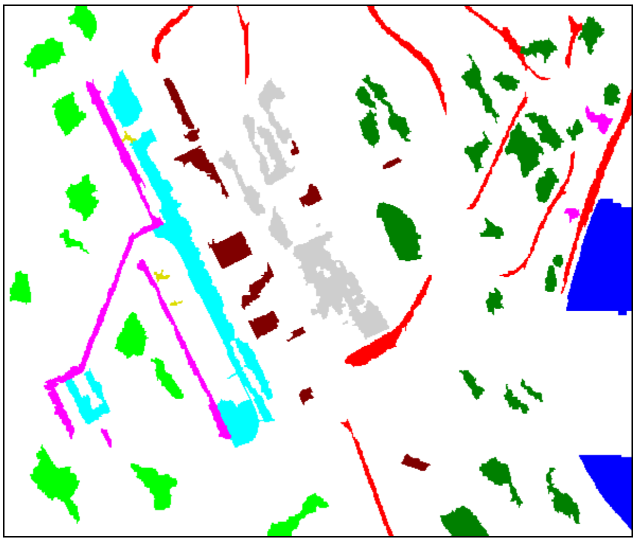

Fig. 6 – Training sites location. See Fig. 6 for legend.

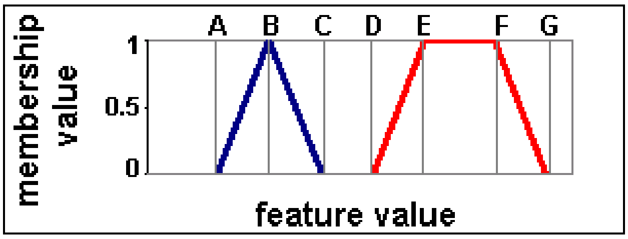

Fig. 7 – Two shapes for fuzzy sets.

Two types of fuzzy sets intersection were analysed: comparison of classes by pairs for each feature and comparison of one class to "others".

5.1 Comparison by pairs

The comparison by pairs of 9 classes with 39 features resulted in 39 matrices of comparison. For each pair, the feature giving the minimum h was preserved. Triangular fuzzy sets were used.

5.2 Comparison of one class to others

When considering one class of interest, training sites belonging to that class were used to compute a triangular fuzzy representing that class. For the object “others”, all training sites not belonging to the class of interest are selected and a trapezoidal fuzzy set is computed. This shape was used because generally, the data distribution was characterized by a platykurtic shape.

6 Experiments

Our feature space optimization (FSO) method was written in the IDL programming language. Once optimal features were defined, classification was performed with eCognition using the nearest neighbor method with the appropriate features.

eCognition offers two methods for nearest neighbor classification: standard NN and NN. With the standard NN approach, the same set of features is applied to all classes while with the NN classifier, a different feature set can be define for each class. Both standard NN and NN were tested. Nearest neighbor classification was chosen because its works on a unique multidimensional decision space thus avoiding to have to chose between a fusion operator. Nevertheless we also present some results of fuzzy sets-based classification in section 6.4.

The eCognition feature space optimization tool was tested in order to have a reference when evaluating the performance of our method. The performance is evaluated thru a confusion matrix by comparing the classification results with the map of Fig. 5. This map was obtained by a standard nearest neighbor classification with eCognition using almost as many training sites as the number of polygons in the scene. From the confusion matrix two measures of performance are computed: the overall accuracy (OA) and the Kappa coefficient (Ka). The OA corresponds to the diagonal (well classified pixels) of a confusion matrix and is weighted by the number of pixel in each class. The kappa coefficient is a measure of confidence that considers the diagonal as well as the errors of commission (classifying a pixel into one class while it belongs to another one) and the errors of omission (classifying a pixel belonging to a class into another class).

6.1 eCognition FSO

The classification of the full set of 39 features with the eCognition standard NN classifier led to the following classification accuracy:

Overall accuracy: 0.621
Kappa measurement: 0.532

The eCognition FSO technique, based on minimum distance, gave a set of 18 features as the most appropriate to classify data. With this 18-features set, the accuracy was:

Overall accuracy: 0.606
Kappa measurement: 0.514

The reason that can explain why the classification results are not better when using the “optimized” feature set is that the global optimization is based on an average minimum distance. The distance can be globally small but locally, between some classes, the distance can be large.

6.2 Comparison by pairs

The comparison of classes by pairs resulted in 39 matrices. For each pairs, the minimum value was kept. The minimum h matrix is shown in Table 1. This table represents the best separability between two classes. A value of 1 corresponds to no separability. When more the one feature were contributing to the minimum h of one pair, only one feature was kept in a way to keep the number of final features as low as possible. By analyzing Table 1 and bands contributing to it, the 11 features of Table 2 were identified as the optimal set for the standard nearest neighbor classifier.

Table 1 – Minimum h matrix for the comparison of the nine classes.

	G	F	W	T	B	P	R	S
helicopter	0	0	0	0	0,241	0	0	0
grass	---	0,197	0	0	0	0	0	0
forest		---	0	0	0	0	0	0,131
water			---	0	0	0	0	0
tarmac				---	0,423	0	0	0,042
building					---	0,562	0,259	0,533
parking						---	0,43	0,68
roads							---	0,527

Table 2 – Set of best features according to the comparison of classes by pairs.

0	Mean blue
4	Ratio red
5	Mean diff. to neighbors red (0)
7	Mean diff. to brighter neighbors red
10	GLCM Correlation all dir. red
11	Mean nir
13	Ratio nir
17	GLCM Contrast all dir. nir
20	Mean ndvi
21	Brightness
27	Density

Concerning the NN classification, a more detailed analysis of fuzzy sets intersections was performed in order the identify appropriate features for each class (Table 3.) Note that when extracting unique features of Table 3 we obtain the global set of Table 2.

Results for standard NN and NN classification are presented in Table 4. This first experiment shows that standard NN and NN classifiers of optimized data set give similar results. In both cases, it is a classification improvement compared to the use of the full data set (section 6.1).

Table 3 – Optimal features, for each class, used in a NN classification.

helicopter	5	7				
grass	11	20	21			
forest	20	21				
water	20					
tarmac	0	4	20			
building	4	5	10	13	27	
parking	7	10	13	27		
roads	5	17	20	27		
soil	0	5	10	11	17	20

Table 4 – Classification accuracies obtained with the comparison by pairs.

Standard nearest neighbor	
Overall accuracy:	0.742
Kappa measurement:	0.675
Nearest neighbor	
Overall accuracy:	0.738
Kappa measurement:	0.670

6.3 Comparison of classes to "others"

When comparing one class to "others" the minimum h matrix is easier to interpret. Table 5 shows the feature contributing to the minimum h and the minimum h itself for each object. We can see that soil has a separability of 0.78 with other classes and that this is given by feature #7. This indicates that we might have some difficulty to differentiate between soil and other objects because of the high value of h.

Here, two experiments for both standard NN and NN classification were performed. In one case classification was computed using the best feature for each object and in the second case, the three best features for each class were used. Results are presented in Table 6. These results show that one feature per object is not sufficient recognize classes and that the use of three features allows to increase the classification accuracy. These results also show that the comparison of classes by pairs leads to slightly better classification accuracy than the global comparison.

Table 5 – Minimum h and its feature contributing to it.

Object	h	Feature	#
helicopter	0	Mean diff. To neighbors red (0)	5
grass	0,374	Ratio nir	13
forest	0,25	Ratio red	4
water	0	Mean NIR	11
tarmac	0,336	Mean NDVI	20
building	0,569	Ratio nir	13
parking	0,562	Ratio red	4
roads	0,402	Density	27
soil	0,788	Mean diff. to brighter neighbors red	7

These results show that one feature per object is not sufficient recognize classes and that the use of three features allows to increase the classification accuracy. These results also show that the comparison of classes by pairs leads to slightly better classification accuracy than the global comparison.

Table 6 – Summary of the results obtained with the comparison of classes to others.

Standard NN classification
• Using the best feature for each object
OA: 0.704
Ka: 0.626
• Using the three best features for each object
OA: 0.730
Ka: 0.657
NN classification
• Using the best feature for each object
OA: 0.510
Ka: 0.409
• Using the three best features for each object
OA: 0.591
Ka: 0.506

The knowledge generalization performed by data grouping in a class "others" results in a loss of small differences between classes. This process might explain the lower performance of the NN classification compared to the standard NN.

6.4 Fuzzy classification

In order to compare nearest neighbor classification performance with fuzzy classification, some fusion results are presented in Table 7. For the conjunctive fusion, the reader can refer to the min-max example of section 4. For other fusion operators we refer to [17]. We can see that the conjunctive fusion operator gives less accurate classification results that other operators. We can also see that the use of optimized feature set provides better results. Finally we can see that fusion provides worst results than nearest neighbor classification. This is caused by the fact the fusion implies that a decision is first taken for each class and then a final decision is taken by keeping the class having the strongest membership value. At the opposite, the nearest neighbor takes a decision in one step.

Table 7 – Some results obtained with fuzzy classification

Using all 39 features
• Conjunctive fusion
OA: 0. 485
Ka: 0. 409
• Quantified adaptive fusion
OA: 0. 615
Ka: 0. 534
Using the 11 features (Table 2, section 6.2.1)
• Conjunctive fusion
OA: 0. 584
Ka: 0. 499
• Adaptive fusion
OA: 0. 626
Ka: 0. 542
• Quantified adaptive fusion
OA: 0. 654
Ka: 0. 571

7 Discussion

During the analysis of classification results, quality assessment was performed by comparing overall accuracy and kappa coefficient. In general, in satellite image classification, overall accuracies in the order of 0.9 is obtained. Here, the OA has never been higher then 0.75, which can be considered as low accuracy. Several points must be considered. First, our classifications are compared with a whole thematic map (Fig. 5). In many studies, classification results are evaluated by the use of test sites resulting in higher accuracy evaluation. Second, our ground truth might not be perfect. Third, our processing was done at the polygonal level implying that image was segmented. If the scale parameter of the segmentation process was determined in order to obtain small polygons to preserve small objects such as the helicopters, this scale was not optimal for buildings delimitations. We can see in Fig. 6 that training sites for buildings do not fit their entire shape (Fig. 5) thus not using the full potential of objects size and shape parameters.

Fig. 8 and Table 8 show the standard nearest neighbor classification results of the 11 features defined by the comparison by pairs (Table 2).

The most significant anticipated classification errors, according to Table 1 should occur between parking and soil (0,68), buildings and parking (0,562) and roads and soil (0,527). This is illustrated relatively well is the confusion matrix of Table 8. This shows that even using 39 features, man-made objects are difficult to discriminate. It also shows that the reliability of the feature space optimization will depend on the initial feature sets.

Among surprising results we can note some confusion (Table 8) between grass and roads and between grass and soil. According to Table 1, the h value in both cases is 0 corresponding to a perfect discrimination. This might be explained by several factors such as the fuzzy sets shape (triangle, trapeze, Gaussian) and the values delimiting these fuzzy sets. The classification method might also be a cause of discrepancy between anticipated and observed results. According to Table 8, 8.1% of the pixels belonging to class "grass" has been classified as "roads". In comparison, with the quantified adaptive fusion, only 1.0% of the pixels grass were classified as roads but more confusion occurred between grass and soils.

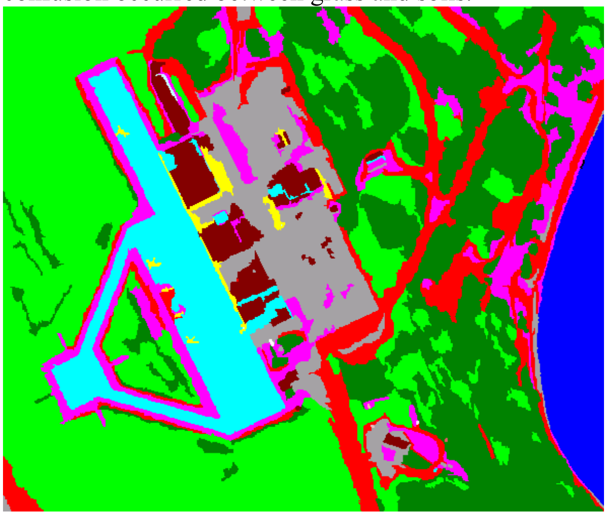

Fig. 8 – Classification of data using the 11 features of Table 2.

Table 8 – Confusion matrix for classification of Fig. 8. Columns: reference data. Rows: classification results. Numbers correspond to pixels frequencies.

	Reference								
	H	G	F	W	T	B	P	R	S
H	240	0	0	0	59	390	1364	0	0
G	0	84215	7849	0	0	0	0	0	0
F	0	26938	43114	0	0	0	143	293	76
W	0	0	0	16780	0	0	0	0	0
T	0	0	0	0	22368	1426	0	0	0
B	0	24	0	0	0	7838	1090	245	0
P	0	189	0	170	4	1485	18833	2598	318
R	0	10280	4657	0	237	6	1709	15118	980
S	0	5265	1231	0	340	292	3294	3441	11691
U	0	0	0	0	0	0	2	0	0

Table 9 – User and producer accuracy derived from Table 8.

	H	G	F	W	T	B	P	R	S
U.A.	1	0,66	0,76	0,99	0,97	0,69	0,71	0,70	0,89
P.A.	0,12	0,91	0,61	1	0,94	0,85	0,80	0,46	0,46

8 Conclusion

We have presented in this study a feature space optimization method based on fuzzy sets intersections. It is a simple technique that use the height of intersection between two fuzzy sets as a measure of feature potential to objects discrimination. Results have demonstrated the improvement of classification accuracy by the use of features optimized with our technique.

If our FSO method shows improvement in the classification accuracy, some point must be kept in mind. First, the results will depend on the classification method and the fuzzy sets shapes. The initial features set is also of great importance. For example, our choice of some initial Haralick texture feature was based on the a priori that many texture features are redundant [18].

Finally another important point to consider is the concordance between the fuzzy sets shape and the histograms from which they were estimated. Using a triangular fuzzy sets for data not normally distributed will lead to inaccurate classification results.

References

[1] Rob J. Decker. Texture analysis and classification of ERS SAR images for map updating of urban areas in the Netherlands. *IEEE. Trans. Geo. Rem. Sen.,* 41(9):1950-1958, 2003.

[2] R.M. Haralick, Statistical and structural approaches to texture, *Proceedings of the IEEE*, 67(5): 786-804, 1979.

[3] B. Solaiman, R. Fiset and F. Cavayas.Automatic road extraction using fuzzy concepts. IGARSS'98, Seattle, USA. 1998.Sunnyvale, CA, 2004.

[4] A.K Shackelford, and C.H. Davis. A hierarchical fuzzy classification approach for high-resolution multispectral data Over urban areas. *IEEE Trans. Geo. Rem. Sen.* 41(9): 1920-1932. 2003.

[5] F. Baret and G. Guyot. Potentials and limits of vegetation indices for LAI and APAR assessment. Rem. Sen. Env. 35: 161-173. 1991

[6] R.O. Duda and P.E. Hart. *Pattern classification and scene analysis*. John Wiley & Sons. NY. 1973.

[7] M. Törmä. Introduction to neural networks and their use in remote sensing. Photogrammetric Journal of Finland. 13(1): 87-102. 1992.

[8] C. Armenakis, F. Leduc, I. Cyr, F. Savopol and F. Cavayas. A comparative analysis of scanned maps and imagery for mapping applications. *ISPRS J. Photo. Rem. Sen*. 57(5-6): 304-314. 2003.

[9] D.A. Landgrebe. Signal Theory Methods in Multispectral Remote Sensing. Wiley-Intersciense. 2003.

[10] W.K. Pratt. *Digital Image Processing*. John Wiley & Sons. NY. 1991

[11] P.M. Narendra and K. Fukunaga. A branch and bound algorithm for feature subset selection. *IEEE Transactions in Computers*. c26(9): 914-922. 1977.

[12] C.J.F. Ter Braak and I.C. Prentice. A theory of gradient analysis. *Advances in Ecological Research*. 18: 272-317. 1988.

[13] B.G. Tabachnick and L.S. Fidell. Using Multivariate Statistics. Ally and Bacon, 4^{th} edition, Needham Heights, MA. 2001.

[14] L.A. Zadeh, Fuzzy sets. *Information and Control*, Vol. 8. pp. 338-353, 1965.

[15] H.-J. Zimmermann. Fuzzy set theory and its applications. 4^{th} edition, *Kluwer Academic Publisher.* 514 p., 2001

[16] B. Solaiman, L.E. Pierce, and F.T. Ulaby. Multisensor data fusion using fuzzy concepts: application to land-cover classification using ERS-1/JERS-1 SAR composites. *IEEE Trans. Geo. Rem. Sen.* 37(3): 1316-1326. 1999.

[17] H. Prade. Possibility theory in information fusion. 3^{rd} *International Conference on Information Fusion 2000*. pp, 5-24. 2000.

[18] A. Baraldi and F. Parmigiani. An investigation of the textural characteristics associated with gray level cooccurrence matrix statistical parameters. *IEEE Trans. Geo. Rem. Sen.* 33(2):293-304. 1995.

[19] D. Dubois and H. Prade. La fusion d'information imprécises. *Traitement du Signal*. 11(6): 447-458. 1995

Appendix 1 – List of the whole initial features set

0	**Mean blue**[1]	*[1]: original features*
1	**Mean green**[1]	*[2]: user computed*
2	**Mean red**[1]	*All other features are computed*
3	Stdev red	*within eCognition*
4	Ratio red	
5	Mean diff. to neighbors red (0)	
6	Mean diff. to darker neighbors red	
7	Mean diff. to brighter neighbors red	
8	GLCM Contrast all dir. red	
9	GLCM Entropy all dir. red	
10	GLCM Correlation all dir. red	
11	**Mean nir**[1]	
12	Stdev nir	
13	Ratio nir	
14	Mean diff. to neighbors nir (0)	
15	Mean diff. to darker neighbors nir	
16	Mean diff. to brighter neighbors nir	
17	GLCM Contrast all dir. nir	
18	GLCM Entropy all dir. nir	
19	GLCM Correlation all dir. nir	
20	**Mean ndvi**[2]	
21	Brightness	
22	Area	
23	Length	
24	Length/width	
25	Compactness	
26	Shape index	
27	Density	
28	Area (including inner polygons)	
29	Perimeter (polygon)	
30	Compactness (polygon)	
31	Number of edges (polygon)	
32	Average length of edges (polygon)	
33	Rectangular angles with edges longer than (polygon) (10)	
34	Length of main line (no cycles) (skeleton)	
35	Length/Width (only main line) (skeleton)	
36	Number of segments (skeleton)	
37	Avrg. area represented by segments (skeleton)	
38	Std.Dev. of area represented by segments (skeleton)	