AUTOMATIC CLASSIFICATION METHODS OF HIGH-RESOLUTION SATELLITE IMAGES: THE PRINCIPAL COMPONENT ANALYSIS APPLIED TO THE SAMPLE TRAINING SET

A.Bernardini, E. S. Malinverni, P. Zingaretti, A. Mancini

DARDUS-Università Politecnica delle Marche, Ancona, IT; - a.bernardini@univpm.it; - e.s.malinverni@univpm.it; - zinga@diiga.univpm.it; - mancini@diiga.univpm.it;

KEY WORDS: Remote Sensing, IKONOS, Classification, Land Cover, Processing, Georeferencing

ABSTRACT:

In Remote Sensing the various bands of multispectral data have not the same relevance in order to identify pixels inside a specific land cover class. The band algebra combines different images in order to construct a new one that has many advantages from the point of view of image understanding or classification (e.g. pseudobands, resulting from the vegetation indices, are used with success for the classification of vegetated areas). The idea of this project was to define new pseudobands through the Principal Component Analysis (PCA) applied to the training sample set of specific classes. We used high resolution IKONOS Multispectral images to test this methodology. PCA was not applied to the whole image, but only to the pixels belonging to a specific class (training sample set). Eigenvectors have a dimension equal to four, like the number of the original bands (Red, Green, Blue and Near Infrared). We selected the Eigenvectors with the highest relevance for a specific class and applied the correspondent orthogonal linear transformation to the whole image in order to obtain the pseudobands containing the relevant information of the chosen class. The same transformation could be applied to the sample training set to obtain a new sample not influenced by the outlier pixels. Pseudobands were segmented by means of a threshold values based on the histograms of the training set Principal Component. A control sample data set was employed to validate the method by means of the Confusion Matrix. The resulting image can be used as mask for the feature segmentation of the selected class.

1. INTRODUCTION

This research is part of a wider project whose main objective is to develop a Corine Land Cover class based automatic classification methodology starting from high resolution Multispectral IKONOS images data set. The images were provided by Regione Marche Institution thanks to a Research Agreement signed together with the Università Politecnica delle Marche jointed departments (DARDUS, DIIGA and DiSASC).

The study case refers to an area belonging to the Ancona Province in Italy (Figure 1). The IKONOS dataset regards an extension of approximately 150 km^2, with both coastal and mountainous aspect, urban and rural landscapes, natural environment, among which the Conero Natural Park has to be mentioned.

According to a hierarchical approach, different phases are applied for land cover classification based on digital image analysis. At the lowest level the analysis is merely based on the pixel spectral values, while moving toward higher levels, the segmentation will carry out the spatial feature patterns. The present project regards the lower level of the analysis. Pseudobands were calculated by means of the linear combination of original bands. We used as linear transformation coefficients the Eigenvectors coming from the sample dataset of a specific land user class. Original images were projected in a space meaningful to discriminate features belonging to the selected class.
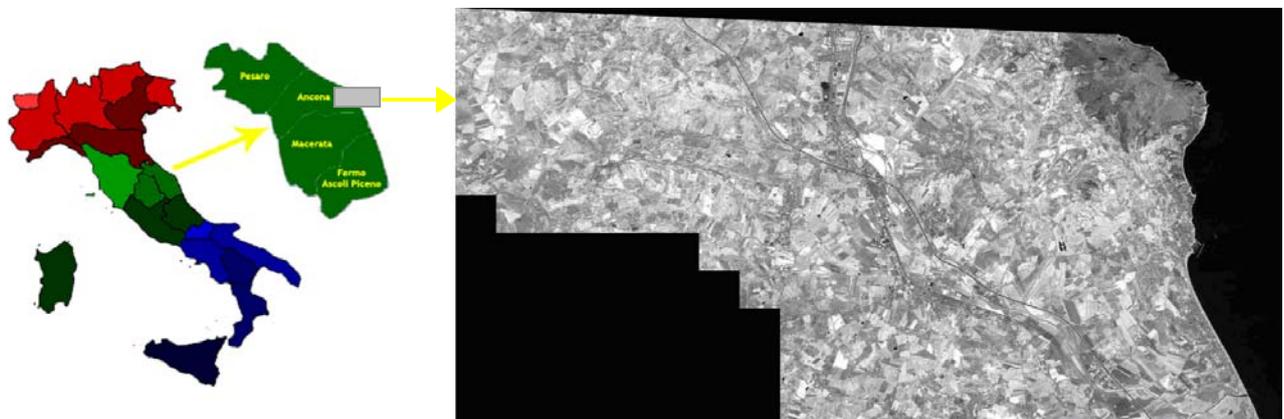

Figure 1. Location of the area study: the surroundings of the Ancona Province, Italy

The work can be divided into four stages (Figure 2) which can be summarized as:

- Pre-processing: the image dataset has been orthorectified in a chosen coordinate system;
- Definition of training and control samples: an exhaustive number of samples, based on a hierarchical classification nomenclature, was the starting point for the PCA and for the assessment of the results;
- Dataset transformation: several transformations, based on different training sample sets, were performed and a thresholding of the resulting bands were carried out. The resulting image can be used as mask in the feature segmentation of the selected classes.
- Accuracy assessment: an evaluation of the accuracy during the thresholding phase was carried out.
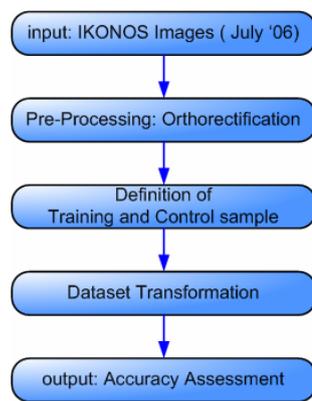

Figure 2. Graphical representation of the 4 stages for the automatic classification of high-resolution satellite images using the PCA

Paper is organized in the following seven sections; Section "Background" introduces the methodological approach used; "Pre-Processing" section provides a description of data orthorectification; "Sample Data set" section points out the necessary procedures to obtain a sufficient number of training samples that are prerequisites for a successful classification. Section "Data set transformation" describes how PCA was used to extract features as road and rail network and associated land. Last two sections, "accuracy assessment" and "conclusions", present and comment the obtained results.

2. BACKGROUND

The Principal Component Analysis is a mathematical method to uncover relationships among many variables and to reduce the amount of data, needed to define the relationships. Applying the Principal Component Analysis to Multispectral images, each band is transformed into a linear combination of orthogonal common components with decreasing variations. Resulting maps, so called "components", carry different information uncorrelated one each others. The components will explain all of the variance contained in the original bands.

Principal component analysis in map algebra is used for several purposes, mainly:

- Transformation of the original bands to obtain new uncorrelated "pseudobands", successively used in the classification process. The greatest amount of the variance in the original dataset can be explained with a lower number of components. The classification process is often improved with the use of these new bands.
- Object recognition: PCA method is used as classification methodology (e.g. face recognition introduced for the first time by Turk and Pentland, 1991).

2.1 Band Transformation

The high correlation of original bands sometimes produces intense computational efforts and generates inefficient results. In Remote Sensing PCA is often used (Gonzalez and Woods, 1993) to compress the information content of n original image channels into a fewer number of transformed Principal Component Images. Principal Components are new uncorrelated bands obtained by linear combination of original data, retaining as maximum as possible the information present on the original data. The transformation of the bands gives rise to a new coordinate system orthogonal to the previous. Among all the PC bands, the first one contains the largest amount of information of the original dataset, while, the last one contains noise. In this way, a classification using the first PCs can be better than one performed by means of the original dataset.

2.2 Object recognition

With this approach, typically used for face recognition (e.g. K.N. Plataniotis 2003), Eigenvectors are calculated on a training set of images, each one representing a sample of objects. In face recognition, commonly the highest Eigenvalues (and associated Eigenvectors) or the most face-like of orthonormal basis vectors, are called Eigenfaces. Bayesian methods for face recognition guarantee the best recognition accuracy rates if compared with PCA, ICA (Independent Component Analysis) and KPCA (Kernel - PCA) (G. Shakhnarovich, B. Moghaddam, 2004]).

The Eigenvectors represent the distinctive features of an image and realise the space of the features: each image can be obtained as linear combination of Eigenvectors and projected on the feature space. In the recognition phase the components of one image on the feature space are calculated and compared with the principal features of other images.

Bajwa, Hyde (Bajwa, Hyde, 2005) applied this methodology to satellite images in the classification of the cloud. In this case, each image of the training sample set represents a cloud; a classification system is defined; each class is represented by a certain number of cloud images in the training sample set. For a successful classification, the classes have to be well separated by a set of features. In training phase, system reads principal features on the cloud images to produce an image space. In testing phase, a new cloud image can be classified by comparing its components in the image space with the components of the training set lying to a class.

3. PRE-PROCESSING

IKONOS Multispectral images were acquired on July 2006, with one 29 degrees solar zenith angle. The IKONOS satellite provides images in panchromatic mode (0.45-0.90 μm) with 1 meter ground resolution and in multispectral mode with 4 meter ground resolution (Table 1).

Band	*Band Centre (μm)*
Band 1 (Blue)	*0.445-0.516 μm*
Band 2 (Green)	*0.506-0.595 μm*
Band 3 (Red)	*0.632-0.698 μm*
Band 4 (Near Infrareds)	*0.757-0.853 μm*

Table 1: Spectral range of IKONOS Multispectral bands
Source: IKONOS Product Guide, Version 1.5, January 2006

The data are available in 11-bit radiometric resolution. They are delivered georeferenced according to the UTM projection and WGS84 Datum. The whole image dataset was orthorectified by means of the Technical Regional Map at the scale 1:10.000, using the rational function model, with the software PCI - Orthoengine. The Ground Control Points (GCPs) were selected on the Map and the DTM derived from the contours lines of the same Map. The geometric correction performed by 15 Control Points identified on the panchromatic image applying a third order rational function, gave an RMS error below 1 pixel (1 meter ground resolution). The radiometric interpolation by means of nearest neighbour re-sampling method preserves the original image values. Next, a sub map of the original image was extracted to define the study area.

4. SAMPLE DATASET

A suitable classification system and a sufficient number of training samples are prerequisites for a successful classification. The system should be informative, exhaustive, and separable (Jensen 1996, Landgrebe 2003).

The first step is the definition of the hierarchical classification structure that is principally based on the user's necessity and spatial resolution of remotely sensed data. On the base of this structure, a significant number of training samples have to be defined.

4.1 Classification structure

The necessity for a standard classification system suggests the use of Corine Land Cover nomenclature, derived from a European Project, led by the European Environmental Agency (EEA) in co-ordination with the member countries. The aim of the project was to define land cover inventories for all European Countries, at the scale 1:100.000, based on a standard methodology and nomenclature, useful for remote sensing applications. The legend has a hierarchical structure on three levels, containing 44 land cover classes grouped into five major categories: Urban Fabric (1), Agriculture Areas (2), Forest and Semi-Natural Areas (3), Wetlands (4), Waterbodies (5). With respect of these structures, the high ground resolution of the IKONOS sensor, suitable for a map at the scale 1:10.000, needs the introduction of others categories of fourth and fifth level. Regarding the Agriculture, the Forest and Semi-Natural Areas, we adopted the extended legend provided by the Italian Agency for Environmental Protection and Technical services (APAT). For Urban Areas we introduced new levels based on the spectral response of different materials.

4.2 Construction of sample set

For each class a sample test set was constructed with labelled polygons. Urban Fabric and Water bodies samples were detected by means of classical photo interpretation techniques or basing on specific surveying campaign for "ground truth" collection. Regarding the categories of Agriculture, Forest and Seminatural Areas, a dedicated dataset research was necessary to minimize errors in land use class assessment. The principal cause of error was the high intra annual land use variation for rotations and short cycle agricultural productions. Approximately 18.000 training pixels were extracted, lying to 27 classes. After a separability analysis (Swain & Davis, 1978), based on histograms of the different bands, some classes were merged into a single class; finally the number of classes was set to 15.

In particular in this project we analysed the Corine Land Cover class called "road and rail network and associated land", in the second level of nomenclature.

We defined some polygons related to road and rail network, associated land (parking areas). For road network and associated land, the sampling activity took into account the different pavement materials (asphalt, gravel, cement). On the base of the separability analysis, we observed that the spectral response of road network, parking areas in asphalt and rail network was very similar. Moreover the same separability analysis, on every land cover types, showed that commercial/industrial buildings with asphalt-type bituminous material covers have very similar spectral response to roads and rail network. So we put together the samples belonging to those categories.

Finally the training set was composed of 34 polygons, of about 2000 pixels. Half of these polygons were used in the training phase and the rest in the assessment phase.

5. DATASET TRANSFORMATION

The analysis was carried out in 4 steps:

1. Construction of training data set. For each class a set of images was made, with the same ground resolution and radiometric information of the original ones, but defined only on the pixels of the data training set.
2. Extraction of the Eigenvectors from the images containing the sample pixels. We processed the four images in order to find four Eingenvectors which make an image space for the selected class.
3. Application of relative transformation on the images. In this phase the data were transformed from the original bands (Red, Green, Blue and Near Infrared) in a new space (the feature space) identified by the first three Eigenvectors. The transformation was made by using the Eigenvectors as coefficients of a linear combination of bands. Each Component, or pseudoband, is a linear combination of the original images.
4. Masking by threshold values. The Pseudobands were segmented by means of threshold values based on the histograms of Principal Components of the training data set.

5.1 Construction of training data set

The analysis was performed with the Open Source software ILWIS (Integrated Land and Water Information System). The software, developed in 1984 by the Netherlands ITC (Institute for Aerospace Survey and Earth Sciences), has been granted open source since 1st July of 2007. It's an integrated Remote Sensing – GIS software that allows to use some internal functions and mathematical calculations in a sequence of instructions organized in script format, to make more than customizable the procedures.

By means of the intersection of sample set polygons with the four IKONOS bands we obtained four images defined only in the pixels corresponding to the training dataset. The Figure 3 shows the histograms of those images.

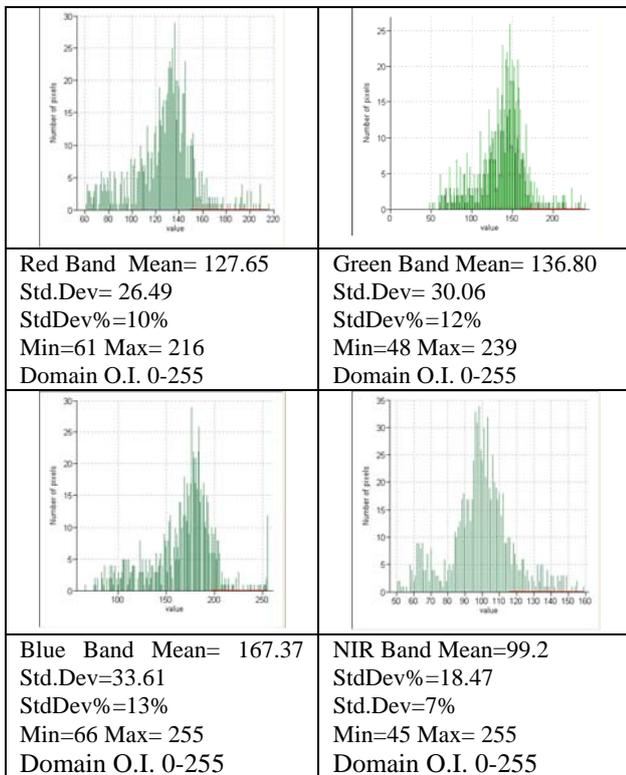|	
Red Band Mean= 127.65 Std.Dev= 26.49 StdDev%=10% Min=61 Max= 216 Domain O.I. 0-255	Green Band Mean= 136.80 Std.Dev= 30.06 StdDev%=12% Min=48 Max= 239 Domain O.I. 0-255
Blue Band Mean= 167.37 Std.Dev=33.61 StdDev%=13% Min=66 Max= 255 Domain O.I. 0-255	NIR Band Mean=99.2 StdDev%=18.47 Std.Dev=7% Min=45 Max= 255 Domain O.I. 0-255

Figure 3. Histograms of the training sample set images on different bands

5.2 Extraction of Eigenvectors from sample images

In this phase we used the PCA tool provided in ILWIS. The input for the Principal Components Analysis consisted of a map list with raster maps of training set images. Starting from this data the covariance matrix and the relative Eigenvectors were calculated. The first Eigenvector explains the 92% of the variance, and the other respectively 6.18%, 1.42% and 0.40%.

5.3 Application of relative transformation on the images

The first three Eigenvectors obtained in the previous phase were used to calculate Principal Components of both training images and original ones (Figure 4).
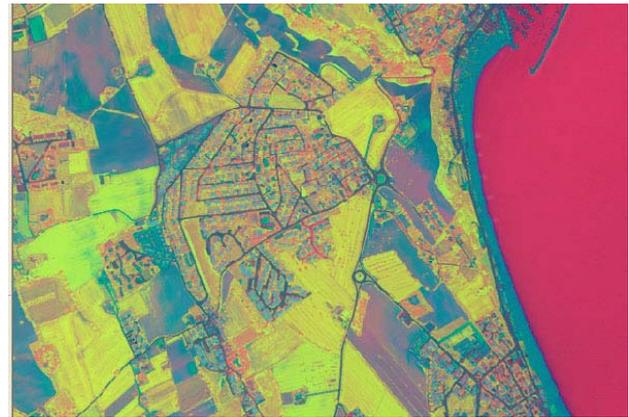

Figure 4. The three Principal Components visualized in false colors

The Figure 5 shows the histograms of the training sample set: we can observe that the variance, expressed as percent of the Domain of definition of the full transformed image, is very small, especially on the second and third component (4%). The transformation produces a reduction in the sample dispersion.

PCA1 Band Mean= 267.60 Std.Dev=53.20 StdDev%=11% Min=112.3 Max= 433.6 Domain T.I. 0-489	PCA2 Band Mean= -29.20 Std.Dev= 13.79 StdDev%=4% Min=-104.6 Max= 2.5 Domain T.I. -240-101

PCA3 Band
Mean= 16.9
Std.Dev= 6.60
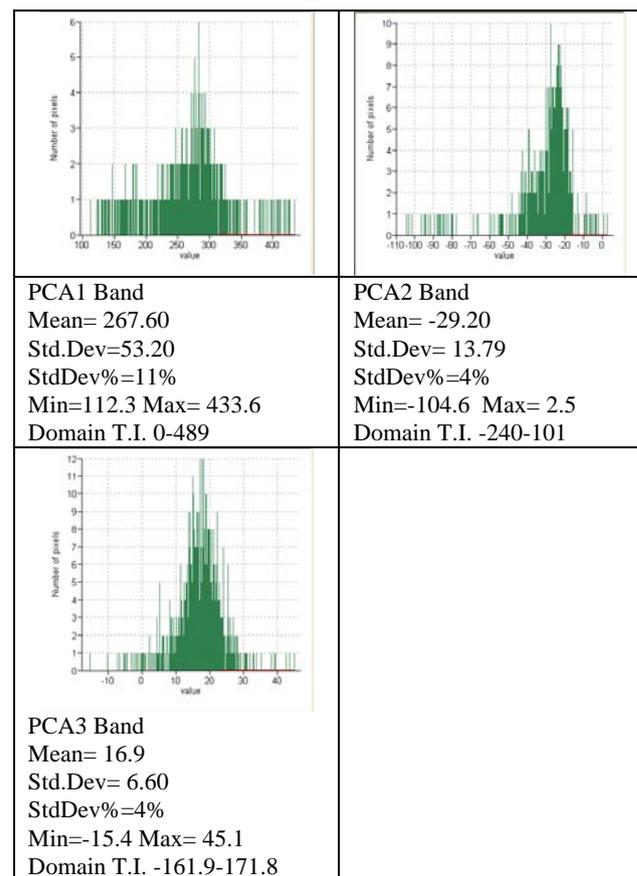StdDev%=4%
Min=-15.4 Max= 45.1
Domain T.I. -161.9-171.8

Figure 5. Histograms of the first three principal components of the training sample set

5.4 Masking by threshold values

The principal idea of this project was to use the Principal Components Domain on training dataset to define the threshold values to apply to the Principal Components of the IKONOS

dataset. The finally masked image was obtained by the intersection of each masked Principal Component.

The Figure 6 shows the result obtained by defining as threshold the minimum and maximum Domain values of the training set components. The 88% of entire image was masked out. We intersected the achieved image with the control dataset, founding that 96% of control dataset was taken out. The remaining 4% consist of mixed pixels inside the commercial/industrial buildings, with asphalt-type bituminous material covers.
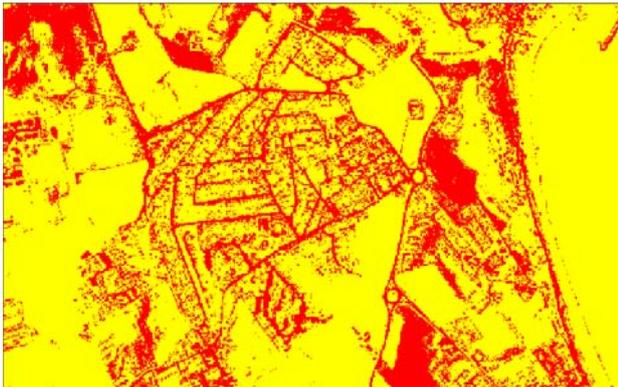

Figure 6. Masked image by the whole range of definition of training set component

By using a narrow interval it is possible to obtain a clearer image: we used both the Mean value ± 2.5Std.Dev and the Mean value ±3 Std.Dev. Figures 7 and 8 point out the results based on these intervals.
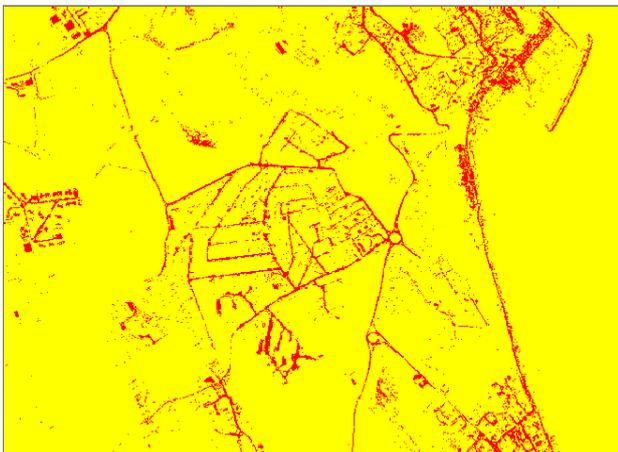

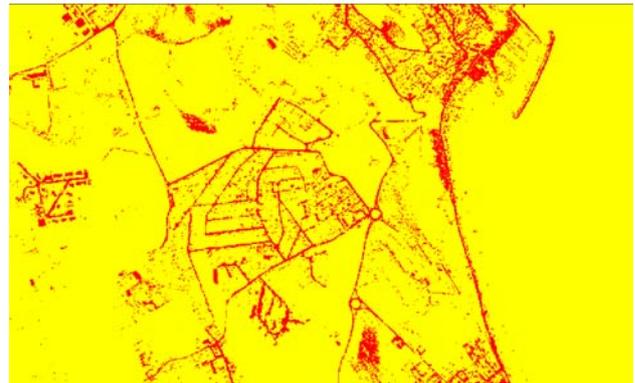Figure 7. Masked image by means of as threshold the Mean value ± 2.5Std.Dev

Figure 8. Masked image by means of as threshold the Mean value ± 3Std.Dev

We can observe that the rail network in last images is clearer than previous one (case of entire range of definition). These binary images can be successful used in road detection with other dedicated feature extraction algorithms.

6. ACCURACY ASSESSMENT

The analysis of the results is based on the Confusion Matrix derived by the Cross processing between the segmented image and the control sample set. As control sample set the overall sample database was used and grouped in two main classes: the first one comprising samples lying to the Road and Rail Network (RRN); the last one including the remaining 14 sample classes (Oth). Furthermore the Confusion Matrix was compared with each one resulting from Maximum Likelihood classification performed by the ENVI software.

	RRN	Oth
RRN	87%	4%
Oth	13%	96%

Table 2. Confusion matrix of PCA thresholding methodology

	RRN	Oth
RRN	67%	0%
Oth	33%	100%

Table 3. Confusion matrix based on Maximum Likelihood classification

The results of PCA thresholding methodology are related to the masked image obtained by the Mean value ± 3 Std.Dev. In this case, the 87% of control dataset was taken out. This percentage increases by using a larger interval of threshold but the resulting image is less clear. In any case this value is bigger than the Maximum Likelihood classification result.

The proposed methodology was applied to Multispectral IKONOS images, with 4 meter Ground Resolution. Figure 9 and 10 represent the overlap of both Panchromatic image and

the Road and Rail Network resulting from the methodology (RRN Image). The RRN image could be improved by means of a noise removal filter and a relinking algorithm to reduce the discontinuity due to the different types of hindering. Furthermore, a procedure has to be tested to use the Panchromatic image, with 1 meter Ground Resolution, in order to refine the geometric shape and location of the road-bed. RNN graph update is the main application. In most of cases, the available graph is not updated. E.g., in Italy during last years, the road topology was influenced by the large scale introduction of roundabouts. Proposed methods can support the graph update process.

Future work will be oriented to use the resulting masked image with other road extraction algorithms.

REFERENCES

Gonzales, R, and Woods, R,1993. Digital Image Processing, Addison-Wesley Publishing Company, Menlo Park, CA pp.148-156

I. S. Bajwa, S. I. Hyder,September 2005. PCA based Image Classification of Single-layered Cloud Types, IEEE, International Conference on Emerging Technologies.

Swain, P. H., & Davis, S. M.,1978. Remote sensing: The quantitative approach. New York: McGraw-Hill, Chapter 3.

M. Turk, A. Pentland, Eigenfaces for Recognition,1991. Journal of Cognitive Neurosicence, Vol. 3, No. 1, pp. 71-86

J. Lu, K.N. Plataniotis, A.N. Venetsanopoulos,January 2003. Face Recognition Using LDA-Based Algorithms, IEEE Trans. on Neural Networks, Vol. 14, No. 1, pp. 195-200

G. Shakhnarovich, B. Moghaddam,December 2004. Face Recognition in Subspaces, Handbook of Face Recognition, Eds. Stan Z. Li and Anil K. Jain, Springer-Verlag.

Figure 9. The Panchromatic IKONOS image

Figure 10. The RRN image overlapped on the Panchromatic image

7. CONCLUSION

This research has presented an automatic classification methods, performed on high-resolution IKONOS images, based on the Principal Component Analysis, applied to the sample training dataset. Starting from the obtained results, the proposed approach allows an efficient segmentation of features of interest (road and rail network and associated land). The quality of results depends by the choice of thresholding mask value. However the accuracy assessment of PCA compared with Maximum Likelihood approach underlines the goodness of proposed method for the selected test class. The achieved road and rail network are useful to graph update the maps at middle scale; unfortunately some extracted areas (e.g. buildings bituminous covers) can influence the network updating process.