Unsupervised Spectral Pattern Recognition for Multispectral Images by means of a Genetic Programming approach

I. De Falco and E. Tarantino
ISPAIM
National Research Council of Italy
Via P. Castellino 111, 80131 Naples, Italy
{i.defalco, e.tarantino}@ispaim.na.cnr.it

A. Della Cioppa
Dept. of Computer Science and Electrical Engineering
University of Salerno
Via Ponte don Melillo 1, 84084 Fisciano (SA), Italy
adellacioppa@unisa.it

Abstract – An innovative approach to spectral pattern recognition for multispectral images based on Genetic Programming is introduced. The problem is faced in terms of unsupervised pixel classification. The system is tested on a multispectral image with 31 spectral bands and 256 $\cdot$ 256 pixels. A good quality clustered output image is obtained.

I. INTRODUCTION

Remote Sensing [1] [2] [3] consists in deriving information about Earth's land and water areas from images taken at a distance. It relies upon measurement of electro-magnetic energy reflected or emitted from the objects of interest at the surface of the Earth. Fields of application include disaster assessment, urban trends monitoring, pollution detection, land use development, water management, erosion assessment, weather forecast, climate changes studies, forest inventarization, and many others [2] [3].

Given an energy source which illuminates target area, Remote Sensing can be accomplished by means of the following steps: record the reflected energy by means of a sensor, transmit recorded information to a receiving station, process data into a digital image and, finally, interpret and analyze this image. This can be made at different wavelengths, resulting in a multispectral digital image.

As concerns the analysis step, a crucial task is feature classification. Classification procedures can be broadly subdivided into *supervised* and *unsupervised* classification. The objective is to assign all pixels in the image to particular classes or themes (e.g. water, coniferous forest, deciduous forest, corn, wheat, etc.). The resulting classified image is comprised of a mosaic of pixels, each of which belongs to a particular theme, and is essentially a thematic "map" of the original image.

A human analyst attempting to classify features in an image uses the elements of visual interpretation to identify homogeneous groups of pixels which represent various features or land cover classes of interest. On the contrary, digital image classification uses the spectral information represented by the digital numbers in one or more spectral bands, and attempts to classify each individual pixel based on this spectral information. This type of classification is termed *spectral pattern recognition*.

Quite recently researchers have started to apply evolutionary techniques to fulfill several tasks related to image understanding. With reference to spectral pattern recognition, nonetheless, at present only little research is reported in literature about supervised pixel classification, and, as far as we know, only one paper deals with unsupervised pixel classification [4].

This paper introduces an innovative approach based on Genetic Programming (GP) [5] to unsupervised pixel classification for multispectral images. Differently from other unsupervised classification approaches, ours is capable of automatically determining the most suitable number of clusters. The paper is organized as follows. Section II describes the problems of spectral pattern recognition and of unsupervised pixel classification. Section III briefly reports on the use of Evolutionary Algorithms in image understanding and particularly in spectral pattern recognition. Section IV focuses on our GP–based approach. In order to assess its feasibility, section V reports preliminary results on the application of our tool to a publicly available multispectral image with 31 bands and $256 \cdot 256$ pixels. The conclusions describe both the positive features and the current limitations of our method. Finally, future works are outlined.

II. SPECTRAL PATTERN RECOGNITION

The aim of spectral pattern recognition is twofold. The first goal is the division of all the multispectral image pixels into clusters, based on statistical features of the pixels themselves. This also results in the identification of the optimal number of clusters.

The second goal is the association of any found cluster with the corresponding material it represents. This can be accomplished based on the fact that the amount of solar radiation which is reflected, absorbed or transmitted by any given material varies with wavelength. This important property of matter makes it possible to identify different substances or classes and separate them by their spectral signatures (spectral curves) identifying uniquely any given material. Huge catalogs exist which contain thousands of spectral signatures, and experts can tell, based on them and on their own experience, which material corresponds to a given spectral curve.

When talking about classes, we need to distinguish between information classes and spectral classes. Information classes are those categories of interest that the analyst is actually trying to identify in the imagery, such as different kinds of crops, different forest types or tree species, different geologic units or rock types, etc. Spectral classes

0-7803-7282-4/02/$10.00 ©2002 IEEE

are groups of pixels that are uniform (or near–similar) with respect to their brightness values in the different spectral channels of the data. The objective is to match the spectral classes in the data to the information classes of interest. Rarely is there a simple one–to–one match between these two types of classes. Rather, unique spectral classes may appear which do not necessarily correspond to any information class of particular use or interest to the analyst. Alternatively, a broad information class (e.g. forest) may contain a number of spectral sub–classes with unique spectral variations. Using the forest example, spectral sub–classes may be due to variations in age, species and density, or perhaps as a result of shadowing or variations in scene illumination. It is the analyst's job to estimate the utility of the different spectral classes and their correspondence to useful information classes.

Common classification procedures can be broken down into two broad subdivisions based on the method used: supervised and unsupervised classification. Basic step in classification is the choice of a decision rule. This can be either parametric or non–parametric. The former are based on known statistical properties, like mean vector and covariances; examples of deriving classification techniques are maximum likelihood, Bayes, Ward [1]. The latter, instead, rely on (non)linear functions or mathematical/geometrical subdivision of the feature space, and result in classification techniques known as minimum distance to means, nearest neighbor, parallelepiped box, and so on [1].

In supervised classification, the task of determining the right number of clusters is left to analysts. In fact, the analyst identifies in the imagery homogeneous representative samples of the different surface cover types (information classes) of interest. These samples are referred to as training areas. The selection of appropriate training areas is based on the analyst's familiarity with the geographical area and his/her knowledge of the actual surface cover types present in the image. Thus, the analyst is "supervising" the categorization of a set of specific classes. The numerical information in all spectral bands for the pixels comprising these areas are used to "train" the system to recognize spectrally similar areas for each class. A special algorithm (of which there are several variations) is used to determine the numerical "signatures" for each training class. Once the algorithm has determined the signatures for each class, each pixel in the image is compared to these signatures and labeled as the class it most closely "resembles" digitally. Thus, in a supervised classification we are first identifying the information classes which are then used to determine the spectral classes which represent them. The main drawback of such a method is that determining a good truth ground may be a very difficult task, since analysts must be sure that any pixel belonging to it is fully representative of the material related to the cluster, and does not contain any feature typical of another material. This may take place because if pixel resolution is not sufficiently high, different elements may be contained in a same pixel. For example, in a pixel representing a $3 \cdot 3$ meters area, we might have a road, a car and tree branches together. This gives origin to the so–called pixel unmixing problem [3].

Unsupervised classification in essence reverses the supervised classification process. Spectral classes are grouped first, based solely on the numerical information in the data, and are then matched by the analyst to information classes when possible. As regards the clusterization phase, classical iterative programs, called *clustering algorithms*, are typically used to determine the natural (statistical) groupings or structures in the data. Firstly, the analyst must specify how many groups or clusters are to be looked for in the data. In addition to specifying the desired number of classes, the analyst may also specify parameters related to the separation distance among the clusters and the variation within each cluster. The final result of this iterative clustering process may result in some clusters that the analyst will want to subsequently combine, or clusters that should be broken down further; each of these situations requires a further application of the clustering algorithm. Once clustering has terminated, matching between the found clusters and the materials they represent must be carried out. This constitutes a very critical task for unsupervised classification, and is accomplished thanks to spectral signatures achieved and to analyst's experience.

III. EVOLUTIONARY ALGORITHMS IN SPECTRAL PATTERN RECOGNITION

There exist many techniques for exploiting the spectral content of multispectral imagery. We can recall here at least Tassel Cap, Atmospherically Resistant Vegetation Index, Normalized Difference Vegetation Index, Principal Component Analysis [1]. Unfortunately these techniques face difficulties when the number of bands increases. It is beyond the scope of this paper to explain into details why it is, suffice it to say here that many of these methods are based on band ratioing, and that, given B bands, there exist $B \cdot (B-1)$ different possible ratios. As the number of bands increases the number of possible combinations becomes rapidly unmanageable. So, approaches based on innovative methods, which can help to automize as much as possible the classification task, are welcome.

Quite recently, researchers have started to take into account approaches based on heuristic techniques like Evolutionary Algorithms. With reference to this latter category, in the following we briefly describe some recent developments. Please note that all of the papers are recent indeed, and the oldest one dates back to 1998, and since then interest in the technique has been increasing.

Viana and Malpica [6] use Genetic Algorithms (GAs) to project a high dimensional space (hyperspectral space) to one with few dimensions in unsupervised classification. Firstly, since the experience shows that bands that are close in the spectrum have redundant information, groups of adjacent bands are taken and a GA is applied in order to obtain the best representative feature for each group, in the sense of maximizing the separability among clusters. Then

0-7803-7282-4/02/$10.00 ©2002 IEEE

the GA is applied again, but this time context information is included in the process.

Yu et alii [7] use a feature selection technique based on GAs to reduce the dimensionality of the feature space and to select features on 224–band Remote Sensing data generated from the NASA/JPL Airborne Visible/InfraRed Imaging Spectrometer (AVIRIS). GAs are combined with Fuzzy Nearest Neighbours Classifiers.

Hung et alii [4] perform unsupervised learning for multispectral image pixel classification by means of a hybrid technique based on GAs and on Differential Competitive Learning performed by an Artificial Neural Network.

Benson at alii [8] examine the evolution of automatic target detection algorithms and their application to the detection of shipping in spaceborne Synthetic Aperture Radar (SAR) imagery. They apply GP which turns out superior to other techniques used in the field.

Stanhope and Daida [9] use GP for both the generation of rules for the target/clutter supervised classification on a set of infrared military vehicles images obtained with SAR, and for the identification of tanks in a set of SAR images. To perform these tasks, previously defined feature sets are generated on the various images, and GP is used to select relevant features and methods of analyzing these features.

Brumby et alii [10] apply a GP algorithm to image feature extraction in Remote Sensing. They aim at finding open waters amidst vegetation. They claim theirs are just preliminary results, and are investigating the GP algorithm parameter space, and the relative importance of crossover and mutation. Their work is based on supervised learning, and on a truth plane.

Fonlupt [11] applies GP to the ocean color problem. This consists in evaluating ocean components concentration (phytoplankton, sediment and yellow substance) from sunlight reflectance values at selected wavelengths in the visible band. He performs supervised learning. Two different sets of experiments are carried out, related to open ocean and coastal waters respectively. GP results to outperform traditional polynomial fits.

Howard and Roberts [12] use a staged GP strategy to automate the task of visual inspection of images aimed to detect objects of interest. They evolve a ship detector for SAR images of the English Channel and a recognizer of motorized vehicles in infrared imagery. Also their method is based on supervised learning.

Rauss et alii [13] describe an initial use of GP as a discovery engine that performs supervised classification from 28–band spectral imagery, aiming at discovering which bands are the most useful for a specific classification task. Their system finds out, for example, that for grass only 18 out of the 28 bands are helpful.

IV. OUR GP APPROACH

In the present paper we introduce an innovative approach to unsupervised spectral pattern recognition based on GP. An important problem when performing unsupervised learning is that we do not know *a priori* which the right number of clusters is. Differently from the classical methods, in our approach such a number is found by the system rather than being set by the human analyst. Therefore, the aim of our automatic system for unsupervised pixel classification is to find both the most suitable number of clusters for the image and the coordinates of any cluster center.

Our method leaves cluster matching task to experts, nonetheless our system allows to provide them with spectral signatures for all clusters we find in the image, thus easing their job.

We have decided to make use of GP because it allows to easily evolve individuals with different numbers of clusters. Also a GA would be able to perform the same task, but it would be more awkward to manage the existence of individuals with different numbers of clusters (thus, with different genotype lengths).

Actually, to effectively face the problem at hand our approach is based on an adaptation of the canonical GP scheme as explained in the following.

Each individual in the population is a tree: its root contains information about the number of clusters, and as many pointers to cluster information nodes as there are clusters. Any given cluster information node has below it exactly *B* nodes, *B* being the number of frequency bands in the multispectral image. Any such band node contains an integer number. More specifically, the set of node types making up trees is very simple: it consists of the three types *Number_of_clusters*, *Cluster* and *Band*. The root node can only be a *Number_of_clusters*, yielding an integer value n_c in the range $[n_{c_{min}}, n_{c_{max}}]$; this value expresses the number of clusters for the pixel classification represented by the tree. Under a *Number_of_clusters* node only *Cluster* nodes are allowed, and their number is exactly n_c. Any *Cluster* node has exactly *B Band* nodes under it, each of them being a terminal node with a constant integer value and representing one of the coordinates of the cluster center.

This means that any tree has just three levels, and that the structure is strongly constrained. Figure 1 reports the example of a tree. It is the GP's task to find out the most suitable number of clusters and the most adequate coordinates in the hyperspace for any such cluster center.

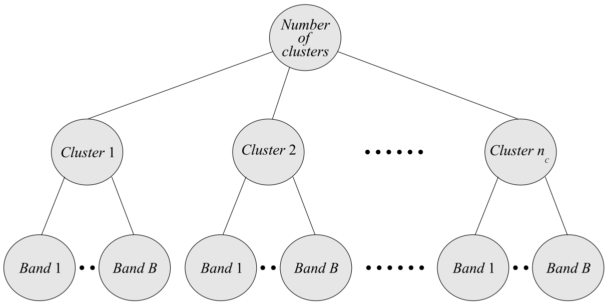

Figure 1: Example of a tree.

The data related to the multispectral image composed by $M \cdot N$ pixels in B bands is stored in an $M \cdot N \cdot B$ matrix

P, often referred to in literature as "data cube". Each element $P(i, j, k)$ of the cube is an integer value in the range $[0, 255]$ expressing reflectance value of pixel (i, j) at k–th wavelength.

Given a tree representing a number of n_c clusters, any pixel is assigned to one and only one among the clusters, namely to the cluster z such that the euclidean distance in B dimensions between the pixel and the center c_z of the cluster z is minimal. In formulae:

$$\text{assign } (i,j) \rightarrow z \text{ if}$$

$$d((i,j), c_z) = \min_{z=1}^{n_c} \sqrt{\sum_{k=1}^{B} (P(i,j,k) - c_z(k))^2}$$

where $c_z(k)$ represents the center of cluster z at k–th wavelength. Let us denote this minimal value with $d_{min}(i, j)$. The fitness ϕ of a tree is then given by:

$$\phi = \sum_{i=1}^{M} \sum_{j=1}^{N} d_{min}(i,j) \qquad (1)$$

and the problem becomes a minimization task.

For any pixel the system records the number corresponding to the cluster it has been assigned to. By doing so, an output image can be drawn which assigns the same color to all pixels belonging to the same cluster.

Our program is based on freeware lil–gp Genetic Programming Software version 1.02 [14], yet strongly modified to best serve our purposes. Therefore, it is based on a set of parameters like number of individuals in the population, number of generations, on several selection methods and on operators of crossover, mutation and reproduction, each with an application probability (breed rate). The software also contains some parameters concerning internal and external crossover probabilities, and internal and external mutations as well. By means of them crossover and mutation points can be chosen more frequently among either internal nodes or leaves. Other lil–gp parameters like initial tree depth range and maximum allowable tree depth are useless in our case, since all legal trees must have a depth of exactly three levels.

To be sure that any new tree has three levels, a transition table has been used for mutations, allowing only legal trees to be generated.

The program gives as output five files containing the output image, the number of clusters and the position of the center for each of them, the number of pixels assigned to any cluster, the spectral signature for each cluster and the evolution of the genetic system.

V. EXPERIMENTAL RESULTS

We have downloaded a publicly available [15] multispectral image with $256 \cdot 256$ pixels and 31 spectral bands ranging from 0.400 up to 0.700 micrometers, each with 0.010 micrometres of width. This means that this image covers the whole visible band, and only it. As a consequence, we can evaluate output image quality even empirically, by looking at it. The image is reported in Figure 2.

As regards the GP parameters, we have chosen a population size of 100 and a maximum number of generations of 2,000. The breed rates are 0.7 for crossover, 0.2 for mutation and 0.1 for reproduction. Selection chosen is tournament with size of five. Internal crossover value is 0.8, while external crossover value is 0.2, meaning crossover being more probable on *Cluster* nodes rather than on *Band* ones. Internal and external mutation probabilities are both set to 0.5. The values for $n_{c_{min}}$ and $n_{c_{max}}$ have been set to 2 and 20, respectively.

A total number of four runs has been carried out up to now by using different seeds for the random number generator. All the evolutions show similar evolutions and similar results in terms of number of clusters and output images.

Figure 3 contains the output image we have obtained at the end of program execution for the best run (in terms of lower final fitness) we have performed at present. More runs would probably get even better quality images. Nonetheless, the image seems reconstructed with very good quality, though only experts can say, based on spectral signatures and on existing catalogs, what any cluster (color in the output image) represents.

The output image contains 17 different clusters, though just three of them represent most of the image with 38.47, 10.55 and 10.52 per cent of the total image respectively. Other clusters represent smaller parts of the image ranging from 0.65 to 4.85 per cent each. A qualitative analysis shows that sky is perfectly separated from the rest of the image. The tree and its branches can be perfectly identified, and so can leaves. The metal plate is clearly defined as well. The bush appears divided into several clusters, depending on the quantity of incident solar radiation, so that directly sunlit leaves look different from shaded ones. The grass is categorized into two different clusters, depending on whether or not direct sunlight reaches the area represented by the pixel. The quality of output image is such that even fallen leaves are visible as small objects in dark colors, and are different from the grass they lie on. Following a mechanism typical in multispectral image analysis, the color associated by our system to any cluster (thus, to any material) has no correspondence to the color by means of which we perceive that material in our visual band. Therefore, as an example, shaded tree leaves are represented in light blue, while the sunlit ones (top left in the image) in dark blue and in violet. Of course, in this case we have perfect knowledge of the scene and we are familiar with all of its components, so our brain can easily reconstruct the meaning of any part. More generally, where multispectral images taken from airplanes or satellites are considered, understanding of the output image might be more difficult, so spectral signatures become really important to tell any area.

An example of multispectral signatures we have achieved in the best run, and related to cluster number 6, is reported in Figure 4. The signature shows a band–pass shape typical of most materials in Remote Sensing, and reaches

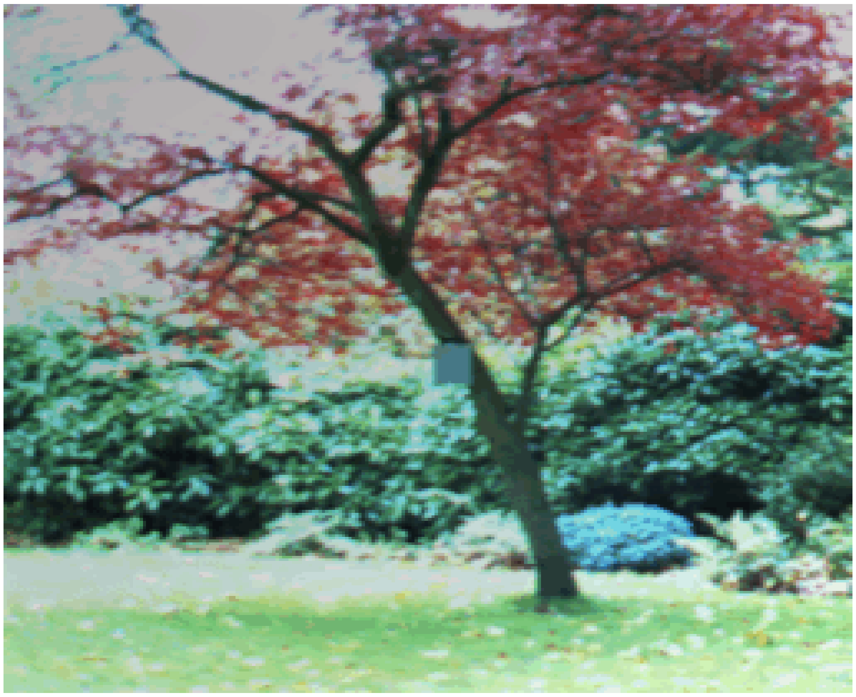

Figure 2: Original image.

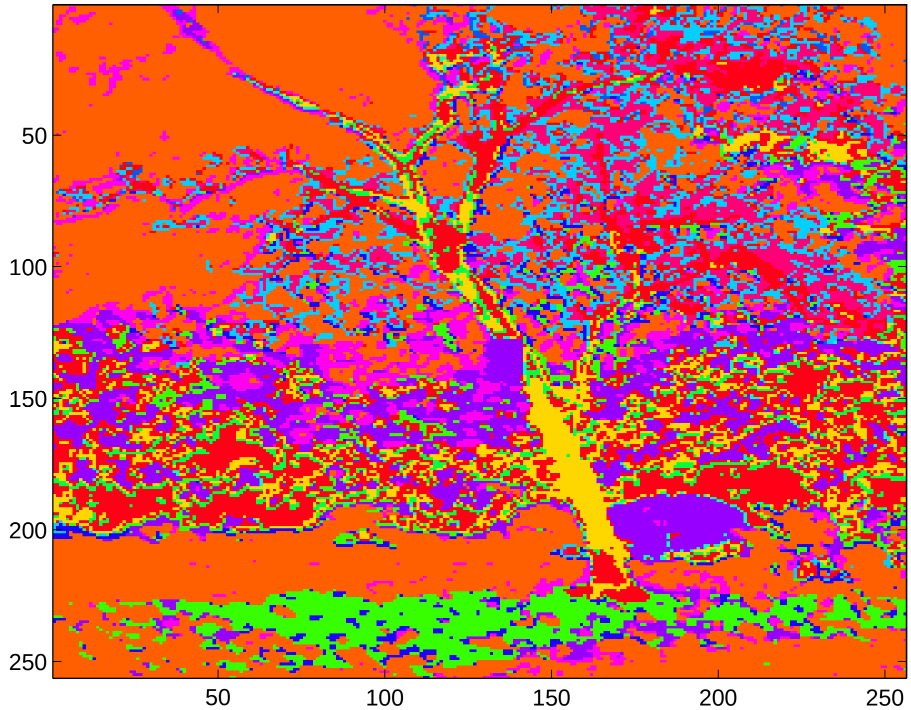

Figure 3: Output image for the best run.

the highest reflectance values in the band between 0.490 and 0.600 micrometers. This band corresponds to green and yellow colors which cover the bands $[0.500, 0.578]$ and $[0.578, 0.592]$, respectively. Therefore cluster 6 might represent sunlit grass.

From the evolutionary point of view, Figure 5 shows the evolution of the best fitness value during the best run performed. As it can be noted, initial generation starts with a best value of about $178,000$. The decrease in fitness values shows a first remarkable quasi–linear phase, until about generation 100, when a value of $120,000$ is reached. Then fitness improvement continues more slowly until generation 800, where a tree with a fitness of about $75,000$ is reached. Since then a new quasi–linear phase starts in which decrease gets slower and slower until end of run, when the best value reached is about $68,000$.

It is interesting to report here that during first generations the system provides us with trees having only few clusters (the best individual in initial generation has just three clusters). As the number of generation increases, solutions consisting of a higher and higher number of clusters are found (four, then five and so on, until a number of seventeen is found and no further changed for many generations). This means that as long as evolution continues, the system reaches higher levels of image details discrimination.

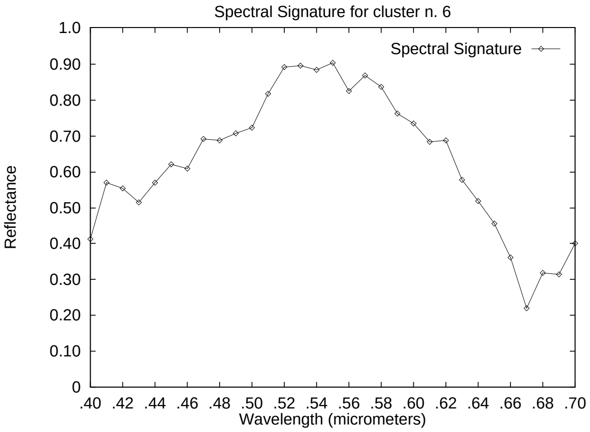

Figure 4: Spectral signature achieved for cluster 6.

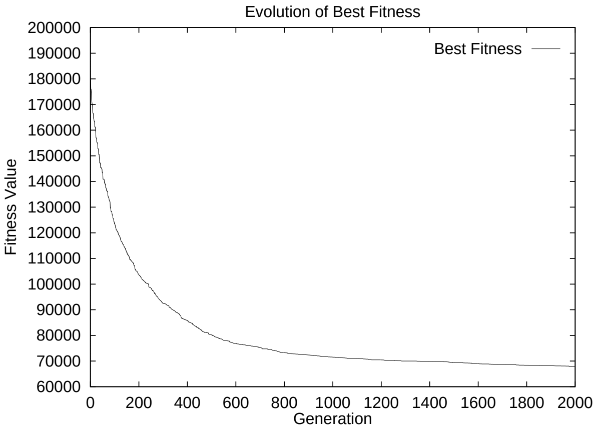

Figure 5: Evolution of best fitness during best run.

VI. CONCLUSIONS AND FUTURE WORK

Aim of the work described in this paper is to test the feasibility of a Genetic–Programming based approach to unsupervised spectral pattern recognition for multispectral images.

The experiments reported here have been conducted by using a publicly available multispectral image with $256 \cdot 256$ pixels and 31 spectral bands.

Achieved results seem positive and encouraging. The output image is of good quality, and material related to any obtained cluster can be determined by means of the yielded spectral signature. Nonetheless we are well aware that these experiments are just preliminary, and are part of an undergoing effort. To obtain a better version of our system and to achieve better results, firstly we need to optimize the tool from the evolutionary point of view. This means to search for the most suitable parameter set. To accomplish this goal, many more runs are needed with the same and with other images.

This leads us to a current drawback of our system: the time needed for one experiment is still high (about two days on a Sun 20 workstation). Therefore, we will do our best in order to reduce execution time. Firstly, code shall be optimized to reduce computing time. Secondly, a parameter set suitable for as many images as possible shall be found. Thirdly, we shall make use of parallel versions of Genetic Programming on Multiple Instruction Multiple Data (MIMD) parallel machines.

From the application point of view, after this tuning phase with publicly available images, we aim to apply our system to hyperspectral (about 90 bands, $512 \cdot 512$ pixels) forestry images coming from European Space Agency (ESA).

Acknowledgements. The authors wish to thank Mr. Vincenzo Mosca for his technical support.

References

[1] H. J. Buiten and J. P. G. W. Clevers, Land Observation by Remote Sensing, Theory and Applications, Gordon and Breach Science Publishers, 1993.

[2] J. B. Campbell, Introduction to Remote Sensing, The Guilford Press, New York, 1987.

[3] T. M. Lillesand and R. W. Kiefer, Remote Sensing and Image Interpretation, John Wiley and Sons Inc., New York, 1994.

[4] C. C. Hung, T. L. Coleman and P. Scheunders, Using Genetic Differential Competitive Learning for Unsupervised Training in Multispectral Image Classification Systems, IEEE International Conference on Systems, Man and Cybernetics, San Diego, CA, pp. 4482–4485, IEEE Press, October 11–14, 1998.

[5] J. R. Koza, Genetic Programming - On the Programming of Computers by means of Natural Selection, MIT Press, 1992.

[6] R. Viana and A. J. Malpica, Genetic Algorithm for Accomplishing Feature Extraction of Hyperspectral Data using Texture Information, Proceedings of SPIE: Image and Signal Processing for Remote Sensing V, Vol. 3871, pp. 367–372, 1999.

[7] S. Yu, S. De Backer and P. Scheunders, Genetic Feature Selection Combined with Composite Fuzzy Nearest Neighbor Classifiers for High-Dimensional Remote Sensing Data, IEEE International Conference on Systems, Man and Cybernetics, Nashville, TN, pp. 1912–1916, IEEE Press, October 8–11, 2000.

[8] K. Benson, D. Booth, J. Cubillo and C. Reeves, Automatic Detection of Ships in Spaceborne SAR Imagery, Genetic and Evolutionary Computation Conference (GECCO), Las Vegas, Nevada, p. 767, Morgan Kaufmann, July 8–12, 2000.

[9] S. A. Stanhope and J. M. Daida, Genetic Programming for Automatic Target Classification and Recognition in Synthetic Aperture Radar Imagery, Evolutionary Programming VII: Proceedings of the Seventh Annual Conference on Evolutionary Programming, Lecture Notes in Computer Science, n. 1147, pp. 693–702, Springer, 1998.

[10] S. P. Brumby, J. Theiler, S. J. Perkins, N. Harvey, J. J. Szymanski, J. J. Bloch and M. Mitchell, Investigation of Image Feature Extraction by a Genetic Algorithm, Proceedings of SPIE: Applications and Science of Neural Networks, Fuzzy Systems and Evolutionary Computation II, Vol. 3812, pp. 24–31, 1999.

[11] C. Fonlupt, "Solving the Ocean Color Problem using a Genetic Programming Approach", *Applied Soft Computing*, Vol. 1:1, pp. 63–72, 2001.

[12] D. Howard and S. C. Roberts, A Staged Genetic Programming Strategy for Image Analysis, Genetic and Evolutionary Computation Conference (GECCO), Orlando, Florida, pp. 1047–1052, Morgan Kaufmann, July 13–17, 1999.

[13] P. J. Rauss, J. M. Daida and S. Chaudhary, Classification of Spectral Imagery using Genetic Programming, Genetic and Evolutionary Computation Conference (GECCO), Las Vegas, Nevada, pp. 726–733, Morgan Kaufmann, July 8–12, 2000.

[14] D. Zongker and W. Punch, lilgp version 1.02, Lansing, Michigan State University, GA Research and Applications Group, 1995. http://isl.cps.msu.edu/GA/software/lil-gp

[15] C. A. Parraca and G. J. Brelstaff, Hyperspectral Dataset, Bristol University, 1995. http://www.crs4.it/gjb/JOSAimages