Feature Extraction from Multiple Data Sources Using Genetic Programming

John J. Szymanski,* Steven P. Brumby, Paul Pope, Damian Eads, Diana Esch-Mosher,
Mark Galassi, Neal R. Harvey, Hersey D.W. McCulloch, Simon J. Perkins, Reid Porter,
James Theiler, A. Cody Young, Jeffrey J. Bloch and Nancy David

Los Alamos National Laboratory
Mail Stop D436, Los Alamos, NM 87545

ABSTRACT

Feature extraction from imagery is an important and long-standing problem in remote sensing. In this paper, we report on work using genetic programming to perform feature extraction simultaneously from multispectral and digital elevation model (DEM) data. We use the GENetic Imagery Exploitation (GENIE) software for this purpose, which produces image-processing software that inherently combines spatial and spectral processing. GENIE is particularly useful in exploratory studies of imagery, such as one often does in combining data from multiple sources. The user trains the software by painting the feature of interest with a simple graphical user interface. GENIE then uses genetic programming techniques to produce an image-processing pipeline. Here, we demonstrate evolution of image processing algorithms that extract a range of land cover features including towns, wildfire burnscars, and forest. We use imagery from the DOE/NNSA Multispectral Thermal Imager (MTI) spacecraft, fused with USGS 1:24000 scale DEM data.

Key words: Multispectral analysis, image processing, evolutionary computation, feature extraction

1. INTRODUCTION AND GOALS

Feature extraction from imagery is an important and long-standing problem in remote sensing. In the case of multispectral imagery, a number of image classification schemes are well known and widely available, but they don't always yield satisfactory results, particularly for classification outside the training data. Typically, a given feature-extraction technique is used on a single image, or perhaps for change detection on a time-series of images. By using multiple data sources for feature extraction, signatures in one sensor can be combined with those in another sensor to produce a signature more easily separated from the background. By signature we mean a spectral band, a mathematical combination of spectral bands, or perhaps the result of spatial processing on a spectral band. Thus, we seek a feature extraction technique that combines the low-level products of sensors, not, for instance, combining the results of supervised classifiers run on multiple sensors. In this paper we investigate a machine-learning approach to feature extraction from multiple data sources.

The particular sample problem for this investigation is land cover classification in the vicinity of Los Alamos, NM. In particular, we will describe our work on separating different land cover classes using data from the US Department of Energy/National Nuclear Security Administration's Multispectral Thermal Imager (MTI) satellite[1,2,3] combined with USGS 1:24k digital elevation model (DEM) data.[4]

The Los Alamos area is of interest not only because we can obtain ground truth easily, but because of the Cerro Grande Fire and rehabilitation efforts post-fire. Between May 6 and May 18, 2000, the Cerro Grande/Los Alamos wildfire burned approximately 43,000 acres (17,500 ha) of forest and 235 residences in the town of Los Alamos. Restoration efforts following the fire were complicated by the large scale of the fire, and by the presence of extensive natural and man-made hazards. These conditions forced a reliance on remote sensing techniques for mapping and classifying the burn region and surrounding vegetation. During and after the fire, remote-sensing data was acquired from a variety of aircraft- and satellite-based sensors, including Landsat 7, MTI, AVIRIS, and others. Data from these sensors are used to evaluate the impact of the fire and begin to monitor the rehabilitation of the ecosystem.

*Correspondence: Email: szymanski@lanl.gov; Telephone: 505-665-9371; Fax: 505-665-0854

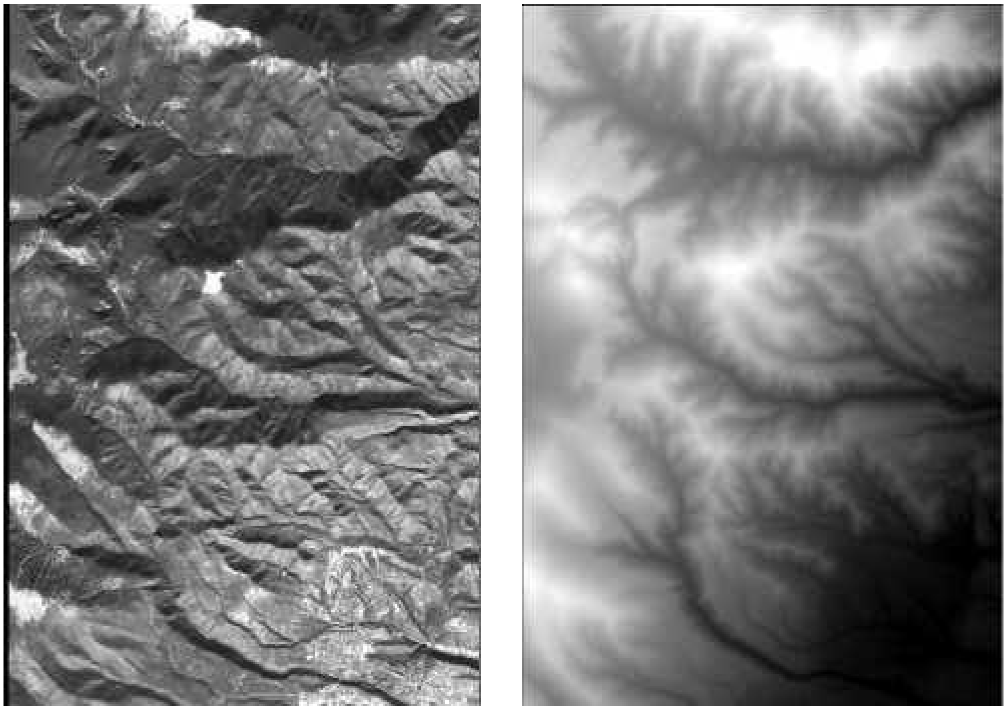

Figure 1. Region of interest. *The left panel shows the part of the Jemez Mountains Northwest of the town of Los Alamos. This is a grayscale representation of a color-infrared slice through the Multispectral Thermal Imager (MTI) image cube (using MTI bands D-C-B). Solar angle is such that north-facing slopes appear dark. The right panel shows the matching DEM data, which is co-registered and appended to the multispectral imagery to form our fused datacube. In the DEM, brighter pixels correspond to higher elevations.*

Beyond classification and mapping of the wildfire burn scars, rehabilitation efforts require up-to-date forest inventories and land cover maps. These can be used to plan rehabilitation efforts, and to estimate remaining forest fuels and hence the risk of further significant wildfires. These mapping products need to be revised on a time scale of years, as destroyed forest gives way to new plantings or as erosion sets in. Land cover map makers have used Landsat TM and ETM+[5] data for many years, and more or less automated algorithms for land cover feature extraction are the subject of an extensive literature.[6,7,8] Such techniques generally require some parameter setting for any given scene, and we are interested in exploring how a machine may learn to set these parameters or find new algorithms for multi-data-source land cover classification.

There are a number of previous efforts that combine multispectral and DEM data. Bucher and Lehmann use high-resolution multispectral data along with hyperspectral data for land cover classification.[9] The DEM data are used both for orthorectification of the multispectral data sets and for differentiating subclasses of vegetation by height. Zhang, Cassells and van Genderen use fused data from a variety of sources for detection and characterization of underground coal fires in China.[10] Their approach makes great use of thermal and multispectral data sources, as well as a DEM, which was used for 3-D visualization of the image data and for deriving depth information about the coal fires. Schistad Solberg, Taxt and Jain explore using Markov Random Fields for multi-source feature extraction, in particular fusing Landsat TM, ERS-1 SAR and GIS for land cover classification.[11] This last reference also gives a good overview of the field.

In previous papers, we described the application of a machine learning technique to the classification of forest fire burn severity[12] and on land cover classification for the entire Jemez Mountain region.[13] Both studies used Landsat 7 ETM+ multispectral imagery without additional data sources. In this work we are also interested in this region's land cover post-fire, with an emphasis on details in the immediate vicinity of the Los Alamos townsite.

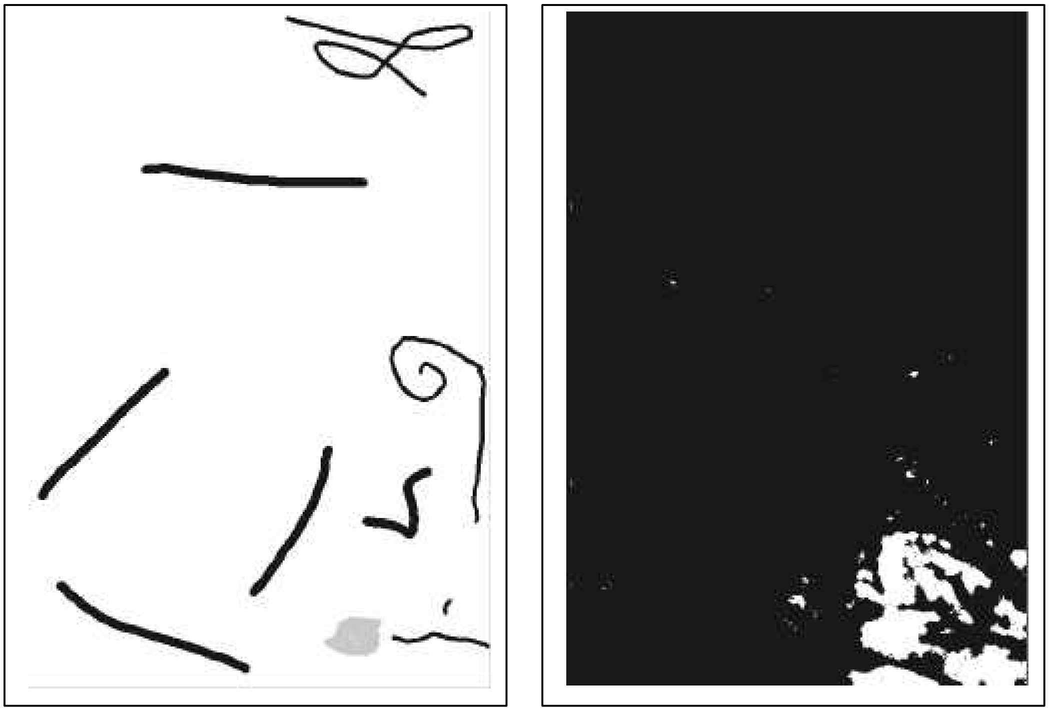

Figure 2. Town/Urban feature. *Left: Training data provided to GENIE. Black pixels define non-town training example pixels, and gray pixels define town example pixels. Right: The GENIE result. Town/Urban was detected in pixels marked in white. Compared to the training data (i.e., only those pixels given a true or false label in left panel), this result achieved a detection rate of 100% and a false alarm rate of 0.16%. Outside of the training pixels, performance is qualitatively good, based on comparison to existing, manual land cover maps and known town boundaries.*

2. TECHNIQUE

Machine learning is an excellent tool for producing feature-extraction algorithms from multiple data sources, for several reasons. First, feature extraction typically involves the setting of multiple parameters, which often don't have a physical basis to determine their values. Second, computers are good at trial-and-error techniques, which are precisely the methods used by many developers of feature-extraction algorithms. Finally, we wish to exploit the serendipitous nature of machine learning – for example: what are the correlations between sensors that we haven't considered, but that result in excellent feature extraction?

In general, should we have a physics-based approach that performs as well as our machine learning algorithms, we would prefer the physics-based algorithm. It is important to note, however, that the algorithms produced in this research are not "black boxes" that are largely impenetrable to the user. Rather the machine learning technique in use here, genetic programming based on genetic algorithms, produces outputs that are interpretable to the user and in fact can be modified by the user.

2.1 GENIE

The tool used in this research, the GENetic Imagery Exploitation (GENIE) software, is showing excellent results in assisted feature extraction tasks. GENIE[14,15,16] is an evolutionary computation (EC) software system that uses a genetic algorithm[17,18,19] (GA) to assemble image-processing algorithms from a collection of low-level ("primitive") image processing operators (e.g., edge detectors, texture measures, spectral operations, and various morphological filters). This system has been shown to be effective in looking for complex terrain features, e.g., golf courses.[20] GENIE can sequentially extract multiple features for the same scene to produce land cover classifications.[21] The implementation details of the GENIE software have been described at length elsewhere,[14,16] so we will only present a brief description of the system below.

GENIE is well suited for feature extraction from multiple data sets because the system does not assume that the input "bands"

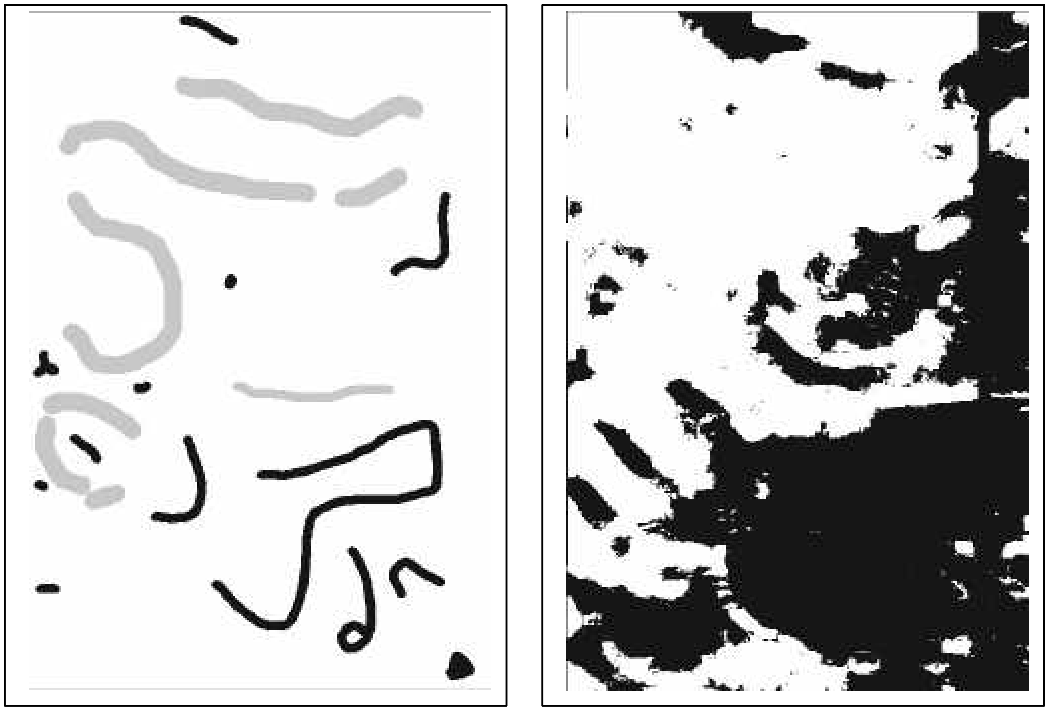

Figure 3. Generic forest feature. *Left: Training data provided to GENIE. Black pixels define non-forest training example pixels, and gray pixels define forest example pixels. Right: The GENIE result. Forest was detected in pixels marked in white. Compared to the training data, this result achieved a detection rate of 99.4% and a false alarm rate of 0.5%. Outside of the training pixels, performance is qualitatively good, based on comparison to existing, manual land cover maps.*

are related to each other in a simple manner. When GENIE runs on a multispectral image, the inputs typically include all the spectral bands of interest, but also may include preprocessed input planes, which may be the output of other classifiers, other genie runs, or hand processing done by the operator. Internal to GENIE are scratch storage planes (the "signature" planes discussed below), which are the result of intermediate computations performed by GENIE. Thus, GENIE does not require all its inputs to be the same type of physical quantity. This, of course, is the case in the present work, where we are combining quite different data sets: multispectral data in radiance units with a DEM in units of meters.

GENIE follows the classic evolutionary paradigm: a population of candidate image-processing algorithms is randomly generated, and the fitness of each individual assessed from its performance in its environment, which for our case is a user-provided training scene. After fitness has been assigned, reproduction with modification of the most fit members of the population follows via the evolutionary operators of selection, crossover, and mutation. The process of fitness evaluation and reproduction with modification is iterated until some stopping condition is satisfied (e.g., a candidate solution with sufficiently high score is found).

The algorithms assembled by GENIE will generally combine spatial and spectral processing, and the system was in fact designed to enable experimentation with spatio-spectral image processing of multi-spectral and hyper-spectral imagery. Each candidate algorithm in the population consists of a fixed-length string of primitive image processing operations. We now briefly describe our method of providing training data and our method for evaluating the fitness of individuals in the population.

The environment for the population consists of one or a number of training scenes. Each training scene contains a raw multi-spectral image data cube, together with a weight plane and a truth plane. The weight plane identifies the pixels to be used in training, and the truth plane locates the features of interest in the training data. Providing sufficient quantities of good training data is crucial to the success of any machine learning technique. In principle, the weight and truth planes may be derived from an actual ground campaign (i.e., collected on the ground at the time the image was taken), may be the result of applying

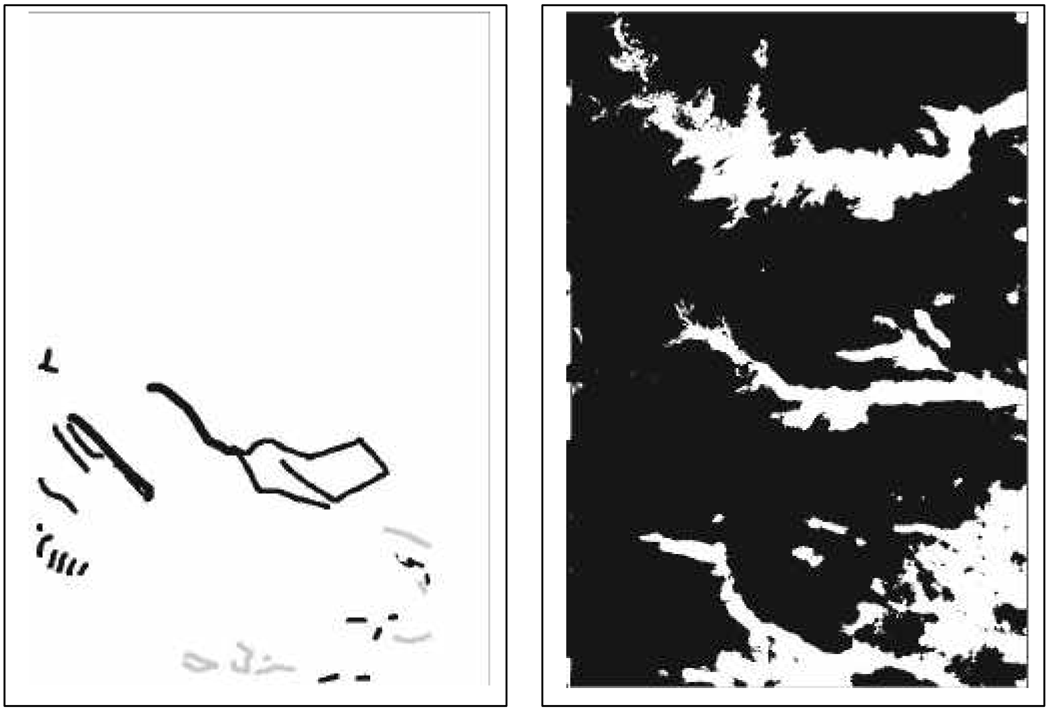

Figure 4. Ponderosa Pine forest feature. *Left: Training data provided to Genie. Black pixels define non-Ponderosa Pine forest training example pixels, and gray pixels define Ponderosa Pine forest example pixels. Right: The Genie result. Ponderosa Pine forest was detected in pixels marked in white. Compared to the training data, this result achieved a detection rate of 99.9% and a false alarm rate of 0.05%. Outside of the training pixels, performance is qualitatively good, based on comparison to existing, manual land cover maps. The algorithm predominantly detects Ponderosa Pine in the medium elevation valleys and canyons bordering on the Jemez Mountains.*

some existing algorithm, and/or may be marked-up by hand using the best judgement of an analyst looking at the data. We have developed a graphical user interface (GUI), called Aladdin, for the manual mark-up of raw imagery. Using Aladdin, the analyst can view a multi-spectral image in a variety of ways, and can mark up training data by painting directly on the image using the mouse. Training data is ternary-valued, with the possible values being “true”, “false”, and “unknown”. True defines areas where the analyst is confident that the feature of interest does exist. False defines areas where the analyst is confident that the feature of interest does not exist. Unknown pixels do not influence the fitness of a candidate algorithm.

Each candidate image-processing algorithm generates a number of intermediate feature planes (or “signature” planes), which are then combined to generate a Boolean-valued mask for the feature of interest. This combination is achieved using a standard supervised classifier (we use the Fisher linear discriminant[22]), and an optimal threshold function.

The fitness of a candidate solution is given by the degree of agreement between the final binary output plane and the training data. This degree of agreement is determined by the Hamming distance between the final binary output of the algorithm and the training data, with only pixels marked as true or false (as recorded in the weight plane) contributing towards the metric. The Hamming distance is then normalized so that a perfect score is 1000.

2.2 The data sets

The data sources used in this work are MTI and the USGS 1:24k DEM, both of which have been described extensively elsewhere.[1-4] The MTI is among many sensors that produced data of the Los Alamos area during or shortly after the fire, and is now supporting ongoing restoration and analysis work, tracking the effects of mitigation efforts and the slow return of vegetation. The MTI image used here was acquired January 13, 2002.

Some preprocessing of the data is needed before Genie can make effective use of them. The MTI instrument team performed calibration and band-to-band registration on their data set. No atmospheric correction is done on the MTI data set. We used the ENVI[23] image processing package to coregister the MTI spectral data and the elevation data contained in the DEM. The coregistration accuracy was less than 3 MTI pixels (i.e. less than 60 meters). This misregistration may be important along the canyon edges, but since the features sought in this work were not expected to depend on small changes in elevation (e.g. along mesa tops and on hillsides), this accuracy was considered sufficient to explore joint MSI/DEM signatures. At this point the coregistered data sets are presented to GENIE for the feature extraction process.

2.3 The Features of Interest

We chose a set of standard land cover classes for which ground truth existed, in the form of an official land cover map[24] for Los Alamos National Laboratory and Los Alamos county. The features we chose to extract were:

- Town/Urban areas
- Wildfire burn scar
- Forest (predominantly Ponderosa Pine, Spruce, Fir, and Aspen)
- Medium elevation Ponderosa pine forest

The topography of the region of interest is quite complex, ranging from the ~10,000 foot peaks of the eastern wall of the heavily forested Jemez Mountains (a dormant volcano), to the ~7,000 foot narrow mesas and steep canyons on which is located the town and Laboratory of Los Alamos. A number of our target features are naturally linked to altitude, e.g., there is an ecological transition region (ecotone) at approximately 8500 feet separating medium altitude forest dominated by Ponderosa Pine, from high altitude mixed conifer forest (which includes Douglas Fir, White Fir, and Spruce), while the town is located on a group of almost constant elevation mesa. Thus, we expect interesting complementary information in the multispectral imagery and digital elevation data sets.

3. RESULTS

For each feature described above, a small amount of training data was marked up by hand using a combination of existing land cover maps, first-hand knowledge of the region of interest, and photo-interpretation of the multispectral imagery. A more extensive mark-up of the scene was also prepared to act as out-of-sample test data for each feature. Figure 1 shows the region of interest, and Figure 2 gives an example of the training mark-up and the GENIE result for the town/urban feature. This mark-up was accomplished using the ALADDIN user interface described above. Each feature is then extracted, one by one, by GENIE in separate processing run on a standard Linux/Intel workstation. Each run required approximately 1 hour of wall clock time. The in-sample (training) and out-of-sample (testing) results for detection rate and false alarm rate for each feature are shown in Table 1.

	In-sample Performance		**Out-of-sample Performance**	
Feature	**Detection Rate**	**False Alarm Rate**	**Detection Rate**	**False Alarm Rate**
Town	100%	0.16%	78.2%	0.2%
Wildfire Burnscar	100%	0.09%	90.25%	2.2%
Forest	99.4%	0.5%	96.6%	2.3%
Ponderosa Pine	99.9%	0.05%	94.4%	14.7%
Ponderosa Pine without DEM	98.8%	3.70%	83.96%	26.8%

Table 1. In-sample and Out-of-sample results for feature extraction using GENIE .

As an example of the individual results, Figure 3 shows training and the GENIE result for the generic forest feature, and Figure 4 shows training data and the GENIE result for Ponderosa Pine. In each case, the qualitative performance of the algorithm compared to the benchmark manual land cover map is good, and gives confidence that the system is learning valid signatures as opposed to simply over-training on the training data. Another paper in this session discusses using GENIE to extract multiple classes simultaneously from a single data type.[25]

The evolved feature extractor with the poorest out-of-sample performance was the town/urban feature extractor. In this case, it appeared that GENIE was only given training data for built-up areas in the town center, and experienced difficulty when tested on urban plus suburban areas. This is understandable, as a large fraction of the suburban component of Los Alamos township is permeated by full-grown trees that fill a substantial aerial fraction when viewed from overhead.

As a test that the system is benefiting from the inclusion of the DEM data, we re-ran the Ponderosa Pine finder problem with the same training data (Fig. 4), but now only presented GENIE with the MTI multispectral imagery. After an equivalent period of training, the performance of the best evolved algorithm (see Table 1) was somewhat less than that of the best algorithm evolved using MSI plus DEM, but the performance outside the training area was noticeably worse, with a substantial decrease in detection rate and a substantial increase in the false alarm rate. In particular, the algorithm trained without access to the DEM data confused Ponderosa Pine forest with high altitude mixed conifer forest throughout the scene.

4. DISCUSSION AND SUMMARY

We have demonstrated evolution of algorithms on a data set consisting of multispectral (visible to thermal) imagery fused with a digital elevation model (DEM). In seeking to evolve algorithms to extract a range of land cover features, including town/urban, forest, and wildfire burnscars, we find that the system was able to exploit successfully the heterogeneous dataset, and continue to perform well outside the training area. We also demonstrated a case where the same evolutionary system trained to find a particular type of forest, Ponderosa Pine, without the DEM data, had difficulty separating the medium elevation Ponderosa Pine forest from the high elevation mixed conifer forest. We find these results encouraging for future efforts of discovering multi-instrument signatures of land cover features.

ACKNOWLEDGEMENTS

We wish to acknowledge our colleagues on the MTI program from Los Alamos and Sandia National Laboratories and the Savannah River Technology Center. We would like to thank Steve Koch, Leslie Hansen, and Randy Balice of LANL's Ecology Group for access to ground truth data and land cover maps used to prepare our training and testing data. This work was supported by the Department of Energy and Department of Defense. We wish to specifically thank Mr. Robert J. Alexander, without whom none of this work would be possible.

REFERENCES

1. W.R. Bell and P.G. Weber, "Multispectral Thermal Imager – Overview," Proc. SPIE **4381**, 173-183 (2001).
2. M.L. Decker, R. Kay, "Multispectral thermal imager satellite hardware status, tasking, and operations," Proc. SPIE **4381**, 184-194 (2001).
3. J. J. Szymanski, W. Atkins, L. Balick, C. C. Borel, W. B. Clodius, W. Christensen, A. B. Davis, J. C. Echohawk, A. Galbraith, K. Hirsch, J. B. Krone, C. Little,P. Mclachlan, A. Morrison, K. Pollock, P. Pope, C. Novak, K. Ramsey, E. Riddle, C. Rohde, D. Roussel-Dupré, B. W. Smith, K. Smith, K. Starkovich, J. Theiler, and P. G. Weber. "MTI Science, Data Products and Ground Data Processing Overview," Proc SPIE **4381**, 195-203 (2001).
4. USGS DEM: http://edcwww.cr.usgs.gov/glis/hyper/guide/1_dgr_dem
5. Landsat TM and ETM+ are described on the U.S. Geological Survey (USGS) web site http://landsat7.usgs.gov
6. R.A. Schowengerdt, *Remote Sensin*g, 2nd ed., Academic, San Diego (1997).
7. J. A. Richards and X. Jia, *Remote Sensing Digital Image Analysis*, 3rd ed., Springer, Berlin (1999).
8. R. S. Lunetta and C. D. Elvidge (editors), *Remote sensing change detection*, Ann Arbor, Chelsea (1998).
9. T. Bucher and F. Lehmann, "Fusion of HyMap hyperspectral with HRSC-A multispectral and DEM data for geoscientific and environmental applications," Proc. of IGARSS 2000: IEEE 2000 International Geoscience and Remote Sensing Symposium. Taking the Pulse of the Planet: The Role of Remote Sensing in Managing the environment (2000).
10. X.M. Zhang, C.J.S. Cassells and J.L. van Genderen, "Multi-sensor data fusion for the detection of underground coal fires," Geologie en Mijnbouw **77**, 117–127 (1999).
11. A.H. Schistad Solberg, T. Taxt and A.K. Jain, "A markov random field model for classification of multisource satellite imagery," Trans. on Geosience and Remote Sensing **34**, 100 (1996).
12. S. P. Brumby, et al., "Evolving forest fire burn severity classification algorithms for multi-spectral imagery", Proc. SPIE **4381**, 236-245 (2001).

13. S. P. Brumby, J. Theiler, J. J. Bloch, N. R. Harvey, S. Perkins, J. J. Szymanski, and A. C. Young. “Evolving land cover classification algorithms for multi-spectral and multi-temporal imagery,” Proc. SPIE **4480**, 120-129 (2002).
14. S.P. Brumby, J. Theiler, S.J. Perkins, N.R. Harvey, J.J. Szymanski, J.J. Bloch, and M. Mitchell, “Investigation of feature extraction by a genetic algorithm ”, Proc. SPIE **3812**, 24-31 (1999).
15. J. Theiler, N.R. Harvey, S.P. Brumby, J.J. Szymanski, S. Alferink, S.J. Perkins, R. Porter, and J.J. Bloch, “Evolving retrieval algorithms with a genetic programming scheme ”, Proc. SPIE **3753**, 416-425 (1999).
16. N. R. Harvey, J. Theiler, S. P. Brumby, S. Perkins, J. J. Szymanski, J. J. Bloch, R. B. Porter, M. Galassi, and A. C. Young,, "Image Feature Extraction: GENIE vs Conventional Supervised Classification Techniques", accepted by IEEE Transactions on Geoscience and Remote Sensing (2002).
17. J. H. Holland, *Adaptation in Natural and Artificial Systems*, University of Michigan, Ann Arbor (1975).
18. I. Rechenberg, *Evolutionsstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution, Fromman-Holzboog*, Stuttgart (1973).
19. L. Fogel, A. Owens and M. Walsh, *Artificial Intelligence through Simulated Evolution*, Wiley, New York (1966).
20. N. R. Harvey, S. Perkins, S. P. Brumby, J. Theiler, R.B. Porter, A. C. Young, A. K. Varghese, J.J. Szymanski, and J.J. Bloch, “Finding golf courses: The ultra high tech approach”, Proc. Second European Workshop on Evolutionary Computation in Image Analysis and Signal Processing (EvoIASP2000), Edinburgh, UK, 54-64 (2000).
21. S. P. Brumby, et al., “A genetic algorithm for combining new and existing image processing tools for multispectral imagery”, Proc. SPIE **4049**, 480-490 (2000).
22. For example, C.M.Bishop, *Neural Networks for Pattern Recognition*, pp.105 –112, Oxford University (1995).
23. See `http://www.rsinc.com/envi`.
24. S. W. Koch (Ecology Group, Los Alamos National Laboratory), private communication.
25. N. R. Harvey et al, in this volume.