IEEE Conference on Systems of Systems Engineering (IEEE SOSE2007), San Antonio, TX, April 16–18, 2007

Min(d)ing the small details: discovery of critical knowledge through precision manifold learning, and application to onboard decision support

Erzsébet Merényi
Department of Electrical and Computer Engineering
Rice University
Houston, Texas 77005, U.S.A.
Email: erzsebet@rice.edu

Lili Zhang
Department of Electrical and Computer Engineering
and Rice Quantum Institute
Rice University
Houston, Texas 77005, U.S.A.
Email: llzhang@rice.edu

Kadim Tasdemir
Department of Electrical and Computer Engineering
Rice University
Houston, Texas 77005, U.S.A.
Email: tasdemir@rice.edu

***Abstract*—Fast identification of critical information in a changing environment is difficult yet it is key to dynamical decision support, in general. Finding critical information in large and complex data volumes is a challenge real systems, and systems of systems, pose increasingly. Moreover, many of these real systems are desired to operate highly autonomously, using extracted critical information and discovered and distilled knowledge directly, for decisions. Spacecraft or rover navigation based on scientific findings from continuously collected data by onboard computation, is one example. This highlights the importance of the quality of information extraction. The knowledge discovery process must be intelligent enough to produce useful details; reliable; robust; and fast. This paper focuses on the first three of these quality aspects through precision manifold learning, in an onboard decision making scenario of a space mission.**

***Index Terms*—Onboard decision support; Embedded learning; Intelligent data understanding; Self-Organizing Maps; Data mining; High-dimensional data.**

I. Background and objectives

Finding critical information in a dynamically changing environment (changing input data) is imperative for a system's decision making and subsequent response. Finding the critical information fast (real, or near-real time) is essential for dynamical decision support. Identification of key information in large and complex data sets is a challenge we face increasingly in systems that are designed to solve complicated tasks. Many of these real systems are desired to operate in a highly autonomous fashion, using extracted information and discovered, distilled knowledge directly for decision making. Unmanned spacecraft operations, or rover navigation, seeking to return data of high scientific value from planetary missions, are examples of such scenarios.

In a recent book "Intelligence for Space Robotics", discussing requirements of successful space missions, Tunstel *et al.* [1] state: "... challenges [in space robotics] may be overcome by increased intelligent sensing, perception, reasoning and autonomous control engineering"; and "... software based navigation intelligence on-board [such] vehicles plays a major role in their safety and success." Many excellent articles in this book emphasize the importance of autonomous navigation, and discuss engineering aspects of it such as hazard avoidance, pointing precision, high performance, reliable onboard computation, fault tolerance, etc. Some, however, adds [2]: "... the most exciting mission opportunities will not be realized without onboard intelligence ...", "... robotic explorers may pass by innumerable scientifically interesting sites, but without the requisite intelligence to recognize them as such, they are simply bypassed and never seen by planetary scientists." The MER missions are quoted as the state-of-the-art for planetary surface robotics [3] for their spectacular success in mobility intelligence. However, while the MERs were highly capable of hazard avoidance and safe navigation based on perceived terrain properties, their general path was pre-designed and commanded by scientists from Earth, in order to accomplish science goals. The rovers did not have onboard processing of science data with sufficient intelligent understanding, to recognize a rare mineralogy or other scientifically relevant surface features, thus could not have made an autonomous decision to stop and examine it. Today's orbiters do not have this capability either, or even just the capability to alert to an interesting event and send the related data (or data product) to Earth with high priority, for preferential human evaluation and intervention.

Autonomous *science driven* navigation of orbiters or rovers must include (in addition to data acquisition) intelligent onboard understanding of scientific data, and knowledge discovery. The fact that this capability is not yet present in space missions has, at least, two reasons. One is that the era of high-performance, reliable, miniaturized and radiation hardened computing facilities, suitable for autonomous onboard operations, has just recently begun [2]. The other is that data from which interesting, new scientific knowledge can be expected is usually very complex, high dimensional, and voluminous. Sifting through such data to discover events

worthy of changing the course of a rover or an orbiter is a formidable task requiring *intelligent algorithms* that are capable of extracting information from data to the extent the sensing device makes it possible.

Several systems are under development, within the NASA community, with the goal of onboard processing of scientific data. ADaM (Algorithm Development and Mining system, http://datamining.itsc.uah.edu/adam/index.html) and EVE (EnVironMent for onboard processing, http://eve.itsc.uah.edu) at the University of Alabama in Huntsville, are primarily designed for mining terrestrial atmospheric and weather data, and data from Earth's magnetosphere. Both of these systems are characterized by very strong and complex architectures for integrating and pushing huge amounts of data through processing pipelines in a well organized fashion, using fairly standard, conventional clustering and classification algorithms for pattern recognition and data mining [4], [5]. Gilmore *et al.* developed a Back Propagation (BP) neural network system for pattern recognition, trained with a large number of laboratory spectra of carbonate and non-carbonate minerals. The trained BP network can successfully alert for the presence of carbonate minerals under a variety of simulated Martian circumstances, in an autonomous regime, thus can be a candidate for onboard decision support [6]. This system, in contrast to ADaM and EVE, is very focused, and constrained to the detection of carbonates from the 2.0 to 2.4 μm wavelength region of remotely sensed spectral data. Gazis and Roush at NASA Ames proposed a rule based Artificial Intelligence (AI) approach for autonomous identification of carbonates (also focusing on the spectral absorption bands near 2.33 and 2.5 μm). This system has been implemented and field tested with reasonable success [7]. Ramsey *et al.* presented a Bayesian system for carbonate identification from Near-Infrared spectra, which is suitable for onboard application, and claims a higher recognition rate than human experts can produce [8].

These examples illustrate the serious interest in onboard processing of scientific data, but they also indicate that the difficulties of the pattern recognition tasks involved are great and can force limited applications. Systems developed to recognize one specific surface feature from a limited subset of the available data (for lack of capabilities to deal with multiple features from all available data), will not recognize other important species. Systems using conventional algorithms may not be able to extract detailed enough knowledge from complex, high-dimensional data such as collected in space missions, and may miss important events.

II. A CANDIDATE KNOWLEDGE DISCOVERY SYSTEM: HYPEREYE

The above underline the *extreme importance of the capability to fully exploit a given data set, and the quality of the extracted information or discovered knowledge*. To enable the widest possible variety of discoveries, and to provide effective decision support, an onboard data understanding subsystem must have the following properties:

1) It must be intelligent enough to deliver high quality information / knowledge, characterized by
   a) high level and precision of useful detail;
   b) repeatability and reliability;
   c) self-assessment of quality, and feedback to the knowledge extraction engines.

   This requires precise learning of the structure of the acquired, often very high-dimensional, data manifold, finding *all* (often a large number of) natural clusters including rare ones, and categorizing them into known and unknown classes. In other words, it is desirable that the system can perform both unsupervised clustering for novelty detection, and supervised classification for known classes of interest, simultaneously. For clustering, the ability of faithful delineation of all clusters, regardless of the distribution of their size, density, shape, etc., capturing of fine intricate structure in the data, is critical. For supervised classification, precise discrimination among many classes with potentially subtle differences between their feature vectors, is imperative. Methods that can take up these challenges are scarce. Dimensionality, for example, is frequently reduced before clustering or classification to accomodate very rich data to algorithms that cannot handle high dimensionality and complexity. This, however, often results in losing discovery or discrimination potential ([9], [10], [11]).
2) A data understanding subsystem must also be capable of continuous learning and adaptation to new situations, since in a space exploration scenario data are acquired continuously;
3) It must be fast (real or near-real time).

These concepts are represented by HyperEye, our manifold learning environment, which we discuss next.

A. HyperEye as an embedded learning subsystem

HyperEye is a collection of neural and other related algorithms for coordinated "precision" mining of complicated and high-dimensional data spaces. It is envisioned to support onboard decision making as depicted in Figure 1. It is specifically designed for the discovery of rare or novel, surprising surface features from data collected by multi- and hyperspectral imagers, as well as for general surface cover mapping of all detected spectral species. This focus is highly motivated because spectral imagers are present now in virtually every planetary mission. The extremely rich data imaging spectrometers provide enable discrimination among a large number of surface materials. Hyperspectral imagers, in particular, provide the spectral resolution that is sufficient to discriminate any surface mineralogy or mixtures of those.

This paper is concerned with the detail and quality of the extracted information or discovered knowledge, as in points 1) and 2) above. Near real-time speed, as in point 3), will be achieved by massively parallel hardware implementation of the neural processors, including on-chip learning capability. This is a non-trivial task for the types and sizes of neural networks

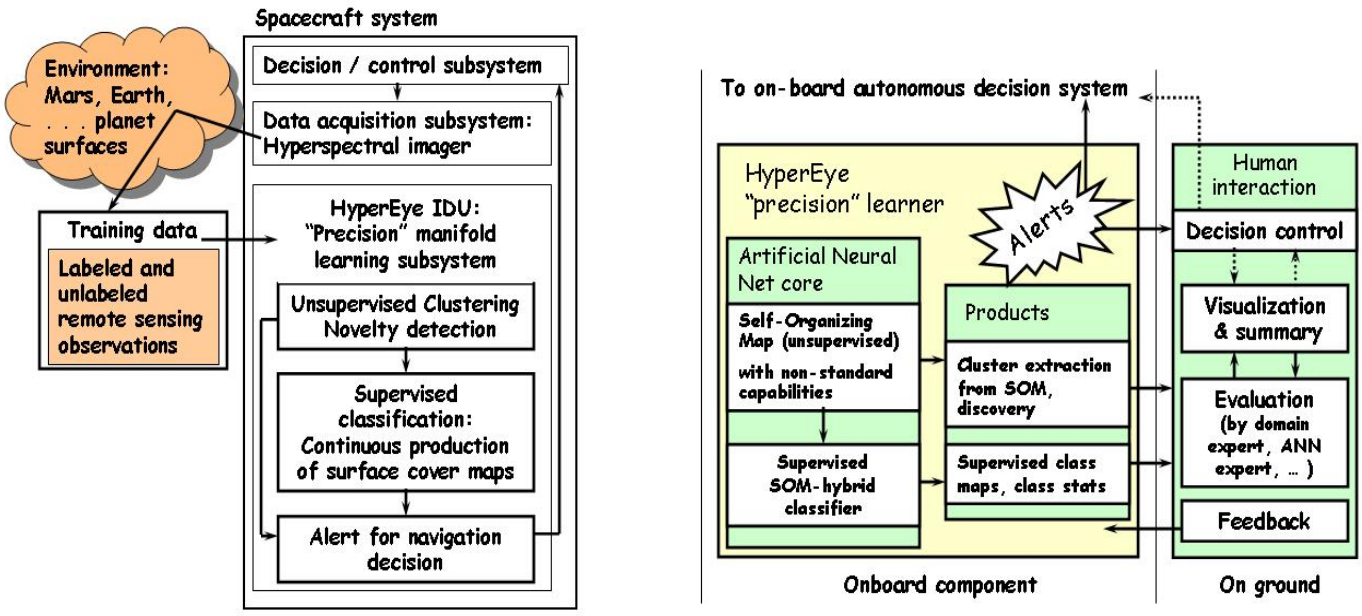

Fig. 1. **Left**: The HyperEye precision manifold learning susbsystem embedded in a spacecraft system, generating scientific knowledge for decision making and navigation control. Knowledge can be extracted using all available data, for maximum discovery potential. Both unsupervised and supervised learning and prediction can go on simultaneously and continuously. Alerts can be generated by either modality, and passed on to the decision and control system. **Right**: The algorithmic components of HyperEye, data products, and their connection to onboard and ground decision making and control.

we use, outside the scope and beyond the current resources of the work involving HyperEye, and will not be discussed here. However, the advances in nanotechnology and interest in real-time neural processing (*e.g.,* [12]) project a hopeful perspective for the not too distant future.

In Figure 1, left, the HyperEye Intelligent Data Understanding (IDU) subsytem, embedded in the spacecraft or rover system, digests and processes data acquired by sensor subsystem(s) from the environment. In this example scenario, the sensor subsystem is a hyperspectral imager, and the environment is a planetary surface. HyperEye has simultaneous unsupervised clustering and supervised classification capabilities as stipulated in point 1) in the previous section. At the heart of these capabilities are sophisticated non-standard neural learning processes that will be discussed below. On this level of operation, the important point is that the IDU subsystem can generate alerts from both unsupervised clustering (upon detection of novel signatures) and from supervised classification (upon finding known interesting species). How the alerts are handled should be defined within the navigation control system.

Figure 1, right, shows details of the HyperEye IDU subsystem: the Artificial Neural Network (ANN) algorithmic core, the main types of data products, and communication of extracted knowledge, in various forms and on various levels of detail, to onboard decision making and/or to humans on ground for feedback. All acquired data can be digested for continuous unsupervised learning of the manifold structure. This is done by a Self-Organizing Map (SOM) and related cluster extractor modules, which have non-standard features and which are central to the sophistication we achieve with HyperEye. We will discuss key details of it below. The learned structure of the data, seen up to the present, is summarized and passed on to a supervised classifier, which utilizes the knowledge of the natural cluster structure of the data for its own learning of labeled data. For example, the underlying known cluster structure helps avoid learning of inconsistent labels, and also helps learning of class boundaries with greater precision than from sparse supervised labels alone. We call this classifier an SOM-hybrid ANN because the SOM is essentially used as a hidden layer in it. Another advantage of the support by the unsupervised clustering is that the supervised classifier can be trained with a much smaller number of training samples than a BP network, and it is much easier to train (does not get easily trapped in local minima). Examples follow below. This help from using unlabeled data is very different from the approach taken by Langrebe *et al.* (see, *e.g.,* [13]), where unlabeled data are gradually folded into the training set of the supervised classifier by labeling them according to the class predictions of the same supervised classifier. While this idea is interesting and has some (idealized) statistical justification it has not been demonstrated for high-dimensional data, and it is unclear how well it would work for data containing many classes with subtle differences.

Labeled data can be provided in advance or during spacecraft operation from known libraries, or generated through on-ground human evaluation of cluster summaries returned by HyperEye. New classes can be added to the supervised classifier as deemed useful. Retraining for new classes is a much lighter load than with a BP network as it does not need to be done from scratch. The neural classifiers in HyperEye, similarly to a BP network, learn a model of the data from

training samples, which provides for more flexible predictions than a fixed rule based AI system can implement and, in general, results in a better success rate. This is especially true for high-dimensional data.

In the rest of this paper we describe some of the key custom features of the SOM(s) we use, as these are the main enablers of the sophistication of HyperEye. We highlight data analysis capabilities through a few examples and in some comparison to other data mining methods.

B. Manifold learning with HyperEye

The core of precision data mining in HyperEye, as stated above, is self-organized manifold learning. Self-Organizing Maps (SOMs), invented by Teuvo Kohonen [14], are intended to mimic the information processing of the cerebral cortex. The algorithm can be briefly described as follows: Let V denote the d-dimensional input data manifold, and A be the rigid, usually 1- or 2-dimensional, SOM grid of Processing Elements (PEs, or neurons), where PEs are indexed by their (potentially multi-dimensional) grid locations $\mathbf{r}$. The weight attached to PE $\mathbf{r} \in A$ is $\mathbf{w_r}$, and is considered a quantization prototype, initially having random values. The SOM learning performs an adaptive vector quantization (VQ) in an iterative process: for any $\mathbf{v} \in V$ input it selects a winner PE $\mathbf{s}$ by

$$\mathbf{s} = \underset{\mathbf{r}}{argmin}||\mathbf{v} - \mathbf{w_r}|| \tag{1}$$

and then updates the weights according to

$$\Delta\mathbf{w_r} = \epsilon h_{\mathbf{rs}}(\mathbf{v} - \mathbf{w_r}) \tag{2}$$

where the *neighborhood function* $h_{\mathbf{rs}}$ defines the weights to be updated, as a function of the grid distance of PE $\mathbf{r}$ from $\mathbf{s}$. $h_{\mathbf{rs}}$ is typically a Gaussian function centered over the winner, but the neighborhood can be defined many different ways.

This quantization differs from other VQ algorithms in two ways. If the learning goes correctly, it finds an optimal placement of the prototypes in data space such that they best approximate the density distribution of the data. In addition, the prototypes will be ordered on the SOM grid according to their similarity relationships. In other words, this is a topological mapping where the neighborhood relationships of the data vectors are preserved on the SOM grid provided that no topology violations occur during learning. (Because of space limitations we refer the reader to basic literature [14] as well as to new research [15], [16] for recognition and remediation of topology violations in SOM learning.) This is a very powerful feature, allowing the detection of contiguous groups of similar prototypes in the SOM grid, which collectively represent clusters of similar data. Cluster boundaries can be identified based on the (dis)similarities (typically Euclidean distances) of the prototypes (not the distances of their SOM grid locations!). SOM clustering does not require an initial guess of the number of clusters (unlike many clustering algorithms), nor does it require any particular initial definition of the quantization prototypes.

SOMs have been popular in the last 20 years and many success stories have been reported. The original Kohonen SOM (KSOM), however, was found suboptimal for high-dimensional data spaces with complex structures. We mention two issues here.

Given an SOM with a fixed size, and K natural clusters in the data (where K is unknown prior to SOM learning), the "real estate" (the number of SOM prototypes) that can be dedicated to the representation of each data cluster is limited. In principle, if the SOM places the prototypes optimally, the *pdf* of the data should be reproduced most faithfully and all clusters (small or large) should have an areal representation proportional to their size. This, however, is not quite so with the KSOM. Theoretical analyses revealed that the KSOM inherently "warps" the grid representation: instead of a linear relationship between the *pdf* of the data, P, and the distribution of the SOM prototypes in data space, Q, which would be characterized by

$$Q(\mathbf{w}) = cP(\mathbf{w})^{\alpha} \tag{3}$$

where $\alpha = 1$, it realizes a functional relationship where $\alpha = 2/3$ in eq (3) [17]. This can lose some clusters when the real estate is tight. For high-volume, complicated data this is always a concern, since the computational cost increases nonlinearly with the size of a 2- (or higher-)dimensional SOM grid. We use a newer variant of the SOM, called *conscience algorithm* [18], which implicitly realizes the $\alpha = 1$ relationship [19]. In addition, as a consequence of the "conscience", one needs only to use an SOM grid neighborhood of a radius of 1 for weight updates in eq (2), which lightens the computational load and accelerates the learning. We also apply other theories to effect a *magnification* of SOM representation areas for rare events, withoung having to know that rare clusters exist in a data set and what they might be. This is done by forcing an $\alpha < 0$ value in eq (3), and thereby enhance the detectability of low-frequency data [19]. An example of this capability is the detection of very rare materials at the Mars Pathfinder landing site, as explained in Figure 2. Full details can be found in [20] and in [21].

The Pathfinder images present a moderate challenge with 8-dimensional spectra (although the complexity is quite high due to calibration differences across the image segments within the SuperPan octants). HyperEye can effectively deal with data of much higher dimensionality. Figure 3 highlights several very (some extremely) small spatial clusters that were discovered from an AVIRIS hyperspectral image of an urban area, using $\sim$200 spectral channels. (This Figure shows approximately half of the spatial area that was clustered.) All discovered features were verified from aerial photographs or by other inquiry. Additional details are given in [11]. This study also contrasts the power of our SOM processing with ISODATA clustering. From Figure 3 one can see that ISODATA confuses cluster assignments in many cases where the SOM cleanly delineates homogeneous surface areas (buildings, golf course, different types of roofs, roads). The mean spectra of all the 35

Fig. 2. Rare surface materials on Mars mapped with HyperEye precision manifold learning from SuperPan data collected by the Imager for Mars Pathfinder. The indicated tiny areas contain a relatively pristine, undifferentiated material termed "black rock" by scientists. This material has a deep 1-μm absorption (olivine or pyroxene) and has been found in very low abundance at the Pathfinder landing site. Our clustering not only found the black rock, but split it into the two subspecies shown in the insets by pale green and hot pink colors. (Please note that both of these colors are unique but to see that among 28 different colors clearly one needs to display the original cluster maps on a high-quality computer screen.) This distinction is justified by the mean spectral shapes of these subclusters (shown in [20]): one has a deeper band centered at 1 μm, the other seems to have its band center beyond 1 μm thus indicating different (undifferentiated) mineralogies. Details can be found in [20]. Note also that many other surface materials have been simultaneously delineated (more than 20 species). Such comprehensive mapping from the Mars Pathfinder data was not done before our work because of the challenges posed by the data.

clusters the SOM discovered, and of the 21 clusters ISODATA produced (shown in [11]) underline significant difference between the two methods. ISODATA not only finds a smaller number of clusters, it does not discover the clusters with the most interesting and unique signatures!

Another important issue we discuss is the extraction of clusters from a learned SOM. From the principle we outlined above, it seems fairly straightforward to delineate cluster boundaries, and in many cases it is so. For high-dimensional data with many natural clusters, especially with widely varying cluster statistics (variable size, density, shapes) and non-linear separability, the detection of cluster boundaries becomes more complicated (*e.g.,* [22]). The representation of cluster (dis)similarities based solely on the weight (prototype) distances in data space (such as in *e.g.,* [23], [24]) is no longer sufficient for confident detection. This problem generated considerable research in recent years, partly because the challenge is intriguing from a manifold learning point of view, but just as importantly because full automation of cluster extraction from SOMs can only be done (in general, for data of high complexity) by overcoming this problem. The problem is worth the effort because the SOM, as shown in the above examples (where we used semi-automated visualization based approaches to extract clusters) does acquire detailed and accurate knowledge about a complicated manifold, in contrast to many other clustering methods including ISODATA. Our challenge is to decipher the SOM's knowledge, and to automate the cluster extraction for autonomous applications.

The structure of a manifold, once quantization prototypes are determined and Voronoi tessellation performed can be described (on the prototype level) by the so-called Delaunay triangulation, which is a graph obtained by connecting the centroids of neighboring Voronoi cells [25]. (This underlines the importance of the optimal placement of the prototypes.) The Delaunay graph has been utilized by many to discover connected and unconnected parts of a manifold (*i.e.,* clusters). With simple data structures this works well. With increased data complexity and noise it becomes very important to portray how *strongly* various parts of the data space are connected. The binary Delaunay graph can show connections caused by a few outliers or by noise between otherwise well separated clusters! Some research started to target this issue recently, to represent (visualize) the connectivity relations of a manifold in order to more precisely delineate clusters. These works, however, are either limited to situations where the SOM prototypes outnumber the data vectors [26], or to data spaces with low dimensions [27], [28]. Clearly, neither solution is sufficient for our goals. We are developing a novel knowledge representation that expresses the manifold connectivities for any data dimension, by showing local data densities overlain on the SOM grid [29]. This has shown promising advantages over existing schemes, for moderate dimensional data sets, and is under further testing and development for high-dimensional data. Our main interest in it is that this knowledge representation will lend itself for automation once we developed it to a satisfactory level of performance. This is in contrast to our current way of extracting clusters semi-automatically from visualizations of prototype distances, which — as the examples show — is successful and detailed, but it would be very hard to capture the human procedure in an algorithm.

We stipulated reliability and self-assessment of quality as essential characteristics of an Inteligent Data Understanding system. While it is fairly straighforward to set quality measures for supervised classifications, such measures are harder to define, but just as important, for unsupervised clustering. To assess the goodness of a clustering without external knowledge (ground truth) is especially important in autonomous environments. The quality of a clustering can be measured, in principle, by assessing how well it matches the natural partitions of a data set. This can provide feedback for an iterative clustering method, to keep improving the clustering until the quality indicator no longer increases. For this purpose, many *cluster validity indices* have been proposed (see, *e.g.,* [30], [31] and references therein), to express to what extent it is true that all data vectors within any cluster are closer to each other than to any data vector in any other cluster. These require no prior knwoledge. In our experience, however, existing indices often misjudge complicated clusterings. This is caused by the metrics they use for within-cluster scatter and for between-clusters separation, which are the main components commonly combined in all validity indices. For example, the popular

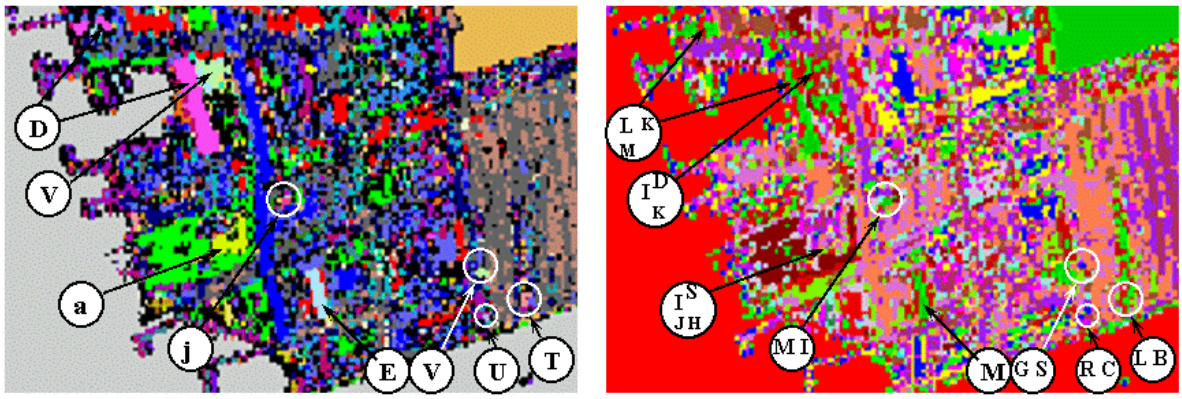

Fig. 3. Details of cluster maps, for a subsection of Ocean City, Maryland, produced from an AVIRIS image using all remaining, ∼190, bands after removal of irrecoverably corrupted bands due to atmospheric attenuation. The spatial resolution is approximately 4 m/pixel. **Left:** SOM clusters. The image contains part of Ocean City, surrounded by sea water (light grey) with board walks extending into the water, small harbors, roads and paved plazas (dark blue and dark grey hues), a large open parking lot at the right (dark grey and mauve colors), beach sand (ocher), vegetation (green colors), and buildings with various roof materials (red, hot pink, light blue, yellow-green). **Right:** ISODATA clusters. Clusters and their colors are different from the SOM map. (ISODATA produced a maximum of 21 clusters even when it was allowed a considerably larger number of clusters. The SOM clustering discovered 35 meaningful clusters.) In both figures, labels and arrows point to the exact same locations. The labels in each figure are given according to the color and label scheme of the respective clustering, and the arrows point to colors that correspond to the labels in the respective circles. The full clustering, complete with corresponding color wedges, can be seen in [11] (downloadable from http://www.ece.rice.edu/∼erzsebet/publist-Merenyi.pdf). Here we point out two things, for comparison. One is that ISODATA confuses clusters where the SOM assigns homogeneous labels. One examples of this is the class D in the SOM map, pointing to a building with a roof that has prominent iron oxide absorptions. ISODATA assigns this building into three different clusters, none of which have signatures with resemblance to iron oxides. (Signatures are not shown here but the full sets are displayed in [11].) The label "a" in the SOM map shows a semi-U shaped building with a very distinct spectral signature. Yet ISODATA fails to delineate it, confusing four different clusters in this building. None of the signatures of those four clusters (I,S,J,H in the ISODATA map) has any similarity to the true spectra at this location. Finally, we point out two tiny spatial clusters: "j" (cherry color), and U (lilac) in the SOM map. "j" is a 6-pixel feature, only occurring here, and at one location within the other image segment we analyzed (not shown here). In both cases, this turned out to be a water tower, as identified from aerial photographs. U points to a sharply delineated 3-pixel feature, with spectral signatures very different from surrounding pixels. This feature was identified as a coast guard lookout tower, from a local map. ISODATA did not discover any of these, or other interesting rare spatial features in spite of their distinctive spectral signatures. The selected features are discussed in detail in [11].

Davies-Bouldin index [30] employs centroid distances for separation measure, which results in favoring spherical clusters. Some indices [32] use data densities, alone or in addition to distances, to better assess clusters of various sizes and shapes. We found a number of widely accepted indices inadequate for assessing our cluster maps, and we are developing new indices designed to provide more faithful measures by taking into account the complicated connectivity relations among high-dimensional clusters of widely variable statistics [31].

Lastly, we give an example of a precise, many-class supervised classification from ∼200-band AVIRIS imagery. The geologic area is Cataract Canyon (in the Grand Canyon), where a landslide hazard study was undertaken as part of a NASA Solid Earth and Natural Hazards Program grant project (PI Victor Baker, U Arizona). The primary purpose of our classification was to map layers in canyon walls with various clay mineralogies as it had been hypothesized that different clays contribute differently to the debris-flooding potential of hill slopes. We show, in Figure 4, half of the resulting class map, and spectral signatures of 15 of the 28 surface cover classes that were mapped. (Readers interested in more specifics including relevant geologic details are referred to [33].) The fine discrimination and sharp delineation of these classes were possible because of the predetermined cluster structure by the SOM in the hidden layer of the supervised classifier.

Theoretical and algorithmic details of the above, with many illustrations, are given in [34], [19] and references therein.

III. Discussion and Future Work

We presented a concept of onboard decision support with HyperEye, as an intelligent data understanding subsytem that extracts critical scientific information from data collected onboard by scientific instruments. By communicating distilled relevant knowledge to onboard (autonomous) and/or on ground (human) decision making systems it is envisioned to contribute to science driven navigation control. In this situation the scope and the quality of the extracted information and knowledge is of paramount importance. We demonstrated some of the current capabilities of HyperEye that we believe can provide smart novelty detection as well as precise detection of a wide variety of known interesting targets, from high-dimensional and complicated data.

While the core functionalities (clustering and classification) of HyperEye produce demonstratedly high quality results, there are outstanding issues to be addressed in order to minimize the need for humans in the decision loop. We discussed two important components of this envisioned onboard IDU subsystem that are incomplete at present: the full automation of cluster extraction from a learned SOM, and the self-assessment of the quality of clustering. With the current readiness, SOM knowledge (including the prototypes and data density counts for each prototype - which is a small amount of data) would be sent to Earth from time to time (or on demand), cluster boundaries extracted semi-automatically by a human analyst, and cluster statistics computed from the

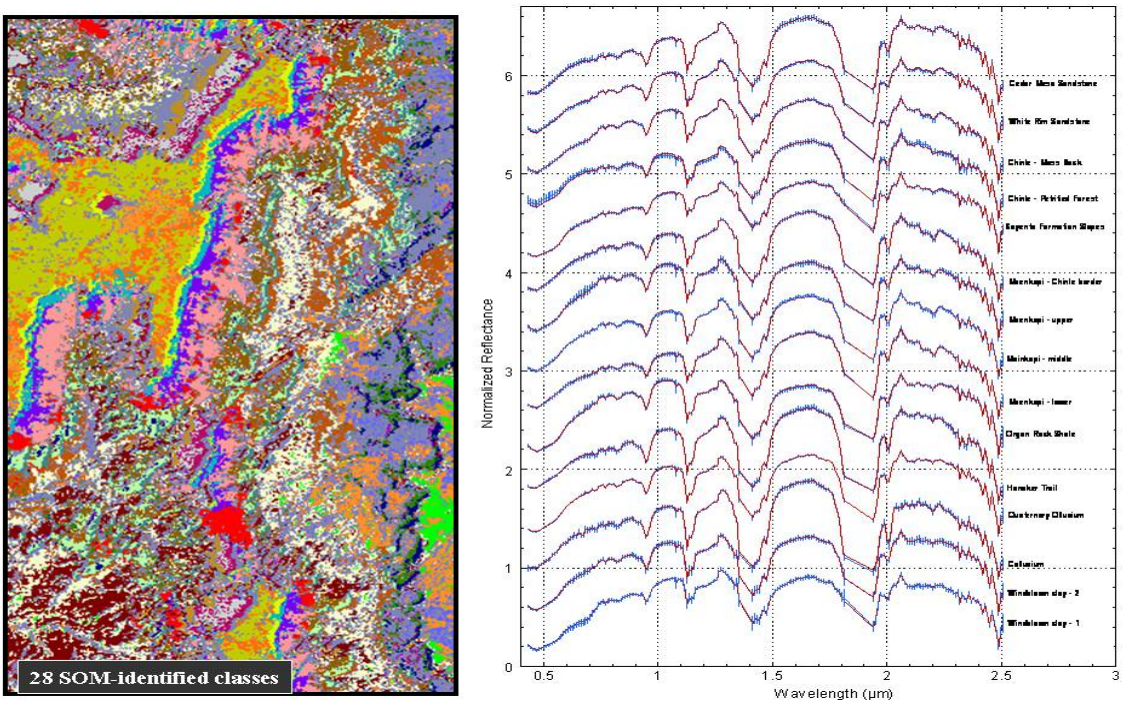

Fig. 4. **Left**: Supervised class map of 28 surface cover types in Cataract Canyon (part of Grand Canyon), Utah, from a 196-band AVIRIS hyperspectral image, using all bands remaining after removing bands with irrecoverable signals due to atmospheric attenuation. Of special interest are a series of layered geologic formations of the Grand Canyon, shown in various colors (blue, turquise, yellow, yellow-green, orange, and others to the right of the blue classes), running down verticaly in the middle, and then continuing with displacements. **Right**: Mean spectra of training sets (blue), and of the predicted classes (red) for 15 of the classes seen on the map at left. The graphs are vertically offset for viewing convenience. The standard deviation of the training classes are shown by vertical bars for each spectral channel. The red mean spectra of the predicted classes are virtually indistinguishable from the training means, indicating tight classification. These spectra represent a situation where precise discrimination of many species was needed, with subtle but meaningful differences in their signatures. Details of this geologic mapping (including the names of the layers, illegible here) are described in [33].

clustered prototypes. This allows novelty detection (since the prototypes of a cluster of data are very similar to the actual data), and decision about appropriate actions. Self-assessment of clustering quality is easy to do at present in an algorithmic sense, but the judgement of available cluster validity indices is unsatisfactory. We are working on remedying this situation.

Interpretation and labeling of newly discovered clusters will need the above human interaction even when cluster extraction will be fully automated. In the long term, this would also be desirable to automate as much as possible, since it can be an extremely time consuming task given the increasing amount of data and knowledge obtained from space missions. One approach would be to create semantic models for planetary data, populate with available data (such as spectral libraries, instrument characteristics, previous analysis results of the same areas) and capture their known relationships. This can help identify the material represented by a "novel" cluster, or ascertain true novelty of it. While a system like this does not exist at present, there are at least partial examples to build on.

One of several related important aspects that we have not discussed in this paper is feature extraction or dimensionality reduction. While we advocate the use of full dimensionality for retention of discovery potential, in situations such as supervised classification, where we know excatly what we are looking for, intelligent feature extraction that takes into account the classification goals, can be extremely beneficial. For this purpose HyperEye has a recently developed neural *relevance learning* module, which performs non-linear feature extraction and has shown significant promise for high-dimensional complex data spaces [35], [36].

We want to emphasize that neural network processing is very slow with sequential computers. Implementation in massively parallel hardware that matches the natural granularity of ANNs, is key to the acceleration of this processing by several orders of magnitude. This is essential for onboard operations, but it is also important for processing large data sets on Earth such as terrestrial archives of remote sensing data. It would, in addition, speed up algorithm development considerably by enabling faster turnaround and testing. High quality clustering of a hyperspectral AVIRIS image can take a couple of days on a regular Sun/Spark workstation. The same could be done under one minute with a massively parallel designated board with currently existing technology [12]. Near-future chips using nanotechnology will be even faster, truly enabling real-time onboard application.

In closing, we add that the methods presented here can be applied directly to similar data such as stacked time series of gene microarrays or spectral images of biological tissues.

They can also be applied to other data with appropriate modifications to ingestion, summarization and housekeeping functions.

ACKNOWLEDGEMENT

The HyperEye concept and algorithm development have been supported by grants NNG05GA94G and NAG5-10432 from the Applied Information Systems Research Program, and by grant NAG5-13294 from the Mars Data Analysis Program of NASA's Science Mission Directorate. This work involves contributions by former graduate students Abha Jain and Major Michael Mendenhall, as well as by former staff member Philip Tracadas and undegraduate ECE major Allen Geer. It also reflects much appreciated stimulation by space and Earth science collaborators.

REFERENCES

[1] E.W. Tunstel, A.M. Howard, and T. Huntsberger, "Robotics challenges for space and planetary robot systems," in *Intelligence for Space Robotics*, A.M. Howard and E.W. Tunstel, Eds., TSI Press Series, pp. 3–20. TSI Press, 2006.

[2] R. Some, "Space computing challenges and future directions," in *Intelligence for Space Robotics*, A.M. Howard and E.W. Tunstel, Eds., TSI Press Series, pp. 21–42. TSI Press, 2006.

[3] M. Maimone, J. Biasiadecki, E. Tunstel, Y. Cheng, and Ch. Leger, "Surface navigation and mobility intelligence on the mars exploration rovers," in *Intelligence for Space Robotics*, A.M. Howard and E.W. Tunstel, Eds., TSI Press Series, pp. 45–69. TSI Press, 2006.

[4] S.J. Graves, "Creating a data mining environment for geosciences," in *Proc. 34th Symposium on the Interface of Computing Science and Statistics, Montreal, Canada*, 17–20 April 2002, vol. http://www.interfacesymposia.org/interface/I02/I2002Proceedings /GravesSara/GravesSara.presentation.ppt.

[5] J. Rushing, R. Ramachandran, U. Nair, S. Graves, R. Welch, and H. Lin, "ADaM: a data mining toolkit for scientists and engineers," *Computers and Geosciences*, vol. 31, pp. 607–618, 2005.

[6] M.S. Gilmore, M.D. Merrill, R. Casta no, B. Bornstein, and J. Greenwood, "Effect of Mars analogue dust deposition on the automated detection of calcite in visible/near-infrared spectra," *Icarus*, vol. 172, pp. 641–646, 2004.

[7] P.R. Gazis and T. Roush, "Autonomous identification of carbonates using near-ir reflectance spectra during the february 1999 marsokhod field tests," *J. Geophys. Res.*, vol. 106, no. E4, pp. 7765–7773, April 25 2001.

[8] Joseph Ramsey, Paul Gazis, Ted Roush, Peter Spirtes, and Clark Glymour, "Automated remote sensing with near infrared reflectance spectra: Carbonate recognition.," *Data Min. Knowl. Discov.*, vol. 6, no. 3, pp. 277–293, 2002.

[9] E. Merényi, "Precision mining of high-dimensional patterns with self-organizing maps: Interpretation of hyperspectral images.," in *Quo Vadis Computational Intelligence: New Trends and Approaches in Computational Intelligence (Studies in Fuzziness and Soft Computing, Vol 54, P. Sincak and J. Vascak Eds.)*. 2000, Physica Verlag.

[10] J. A. Benediktsson J. R. Sveinsson and et al., "Classification of very-high-dimensional data with geological applications," in *Proc. MAC Europe 91*, Lenggries, Germany, 1994, pp. 13–18.

[11] E. Merényi, B. Csató, and K. Taşdemir, "Knowledge discovery in urban environments from fused multi-dimensional imagery," in *Proc. IEEE GRSS/ISPRS Joint Workshop on Remote Sensing and Data Fusion over Urban Areas (URBAN 2007).*, P. Gamba and M. Crawford, Eds., Paris, France, 11–13 April 2007.

[12] M. Porrmann, U. Witkowski, and U. Rückert, "Implementation of self-organizing feature maps in reconfigurable hardware," in *FPGA Implementations of Neural Networks*, A. Omondi and J. Rajapakse, Eds. Springer-Verlag, 2005.

[13] Q. Jackson and D.A. Landgrebe, "An adaptive classifier design for high-dimensional data analysis with a limited training data set," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 12, pp. 2664–2679, December 2001.

[14] T. Kohonen, *Self-Organizing Maps*, Springer-Verlag, Berlin Heidelberg New York, 1997.

[15] Th. Villmann, R. Der, M. Herrmann, and Th. Martinetz, "Topology Preservation in Self–Organizing Feature Maps: Exact Definition and Measurement," *IEEE Transactions on Neural Networks*, vol. 8, no. 2, pp. 256–266, 1997.

[16] L. Zhang and E. Merényi, "Weighted differential topographic function: A refinement of the topographic function," in *Proc. 14th European Symposium on Artificial Neural Networks (ESANN'2006)*, Brussels, Belgium, 2006, pp. 13–18, D facto publications.

[17] H. Ritter, "Asymptotic level density for a class of vector quantization processes," *IEEE Trans. on Neural Networks*, vol. 2, pp. 173–175, 1991.

[18] D. DeSieno, "Adding a conscience to competitive learning," in *Proc. Int'l Conference on Neural Networks (ICNN), July 1988*, New York, 1988, vol. I, pp. I–117–124.

[19] E. Merényi, A. Jain, and T. Villmann, "Explicit magnification control of self-organizing maps for "forbidden" data," *IEEE Trans. on Neural Networks, in press*, 2007.

[20] E. Merényi, W.H. Farrand, and P. Tracadas, "Mapping surface materials on Mars from Mars Pathfinder spectral images with HYPEREYE.," in *Proc. International Conference on Information Technology (ITCC 2004)*, Las Vegas, Nevada, 2004, pp. 607–614, IEEE.

[21] E. Merényi, A. Jain, and W.H. Farrand, "Applications of SOM magnification to data mining," *WSEAS Trans. on Systems*, vol. 3, no. 5, pp. 2122–2128, 2004.

[22] J. Vesanto and E. Alhoniemi, "Clustering of the self-organizing map," *IEEE Transactions on Neural Networks*, vol. 11, no. 3, pp. 586–600, May 2000.

[23] A. Ultsch, "Self-organizing neural networks for visualization and classification.," in *Information and Classification — Concepts, Methods and Applications*, R. Klar O. Opitz, B. Lausen, Ed., pp. 307–313. Springer Verlag, Berlin, 1993.

[24] M.A. Kraaijveld, J. Mao, and A.K. Jain, "A nonlinear projection method based on Kohonen's topology preserving maps," *IEEE Trans. on Neural Networks*, vol. 6, no. 3, pp. 548–559, 1995.

[25] Th. Martinetz and K. Schulten, "Topology representing networks," *Neural Networks*, vol. 7(3), pp. 507–522, 1994.

[26] G. Polzlbauer, A. Rauber, and M. Dittenbach, "Advanced visualization techniques for self-organizing maps with graph-based methods," in *Proc. Intl. Symp. on Neural Networks (ISSN05)*, 2005, pp. 75–80.

[27] M. Aupetit, "Visualizing the trustworthiness of a projection," in *Proc. 14th European Symposium on Artificial Neural Networks, ESANN'2006, Bruges, Belgium*, Bruges, Belgium, 26-28 April 2006, pp. 271–276.

[28] M. Aupetit and T. Catz, "High-dimensional labeled data analysis with topology representing graphs," *Neurocomputing*, vol. 63, pp. 139–169, 2005.

[29] K. Tasdemir and E. Merényi, "Data topology visualization for the Self-Organizing Map," in *Proc. 14th European Symposium on Artificial Neural Networks, ESANN'2006, Bruges, Belgium*, Bruges, Belgium, 26-28 April 2006, pp. 125–130.

[30] J.C. Bezdek and N.R. Pal, "Some new indexes of cluster validity," *IEEE Trans. System, Man and Cybernetics, Part-B*, vol. 28, no. 3, pp. 301–315, 1998.

[31] K. Tasdemir and E. Merényi, "A new cluster validity index for prototype based clustering algorithms based on inter- and intra-cluster density," in *Proc. Int'l Joint Conf. on Neural Networks (IJCNN 2007)*, Orlando, Florida, USA, August 12–17 2007, p. submitted.

[32] M. Halkidi and M. Vazirgiannis, "Clustering validity assessment using multi representatives," in *Proc. of SETN Conference, Thessaloniki, Greece, April*, 2002.

[33] L.Rudd and E. Merényi, "Assessing debris-flow potential by using aviris imagery to map surface materials and stratigraphy in cataract canyon, Utah," in *Proc. 14th AVIRIS Earth Science and Applications Workshop*, R.O. Green, Ed., Pasadena, CA, May 24–27 2005.

[34] T. Villmann, E. Merényi, and B. Hammer, "Neural maps in remote sensing image analysis," *Neural Networks*, vol. 16, pp. 389–403, 2003.

[35] M.J. Mendenhall and E. Merényi, "Generalized relevance learning vector quantization for classification driven feature extraction from hyperspectral data," in *Proc. ASPRS 2006 Annual Conference and Technology Exhibition*, Reno, Nevada, May 5–8 2006, p. 8.

[36] M.J. Mendenhall and E. Merényi, "Relevance-based feature extraction for hyperspectral images," *IEEE Trans. on Neural Networks, under review*, 2007.