# Multi-level Indexing and GIS Enhanced Learning for Satellite Imageries

Krzysztof Koperski
Data Analysis Products Division of Mathsoft, Inc.
1700 Westlake Ave. N, Suite 500,
Seattle, WA, 98109-3044 USA
(206) 283-8802 Ext. 243
krisk@statsci.com

Giovanni B. Marchisio
Data Analysis Products Division of Mathsoft, Inc.
1700 Westlake Ave. N, Suite 500,
Seattle, WA, 98109-3044 USA
(206) 283-8802 Ext. 280
giovanni@statsci.com

## ABSTRACT

Satellite technology produces data at an enormous rate. Most of the database research on the analysis of remotely sensed images concentrated on data retrieval and simple queries that involved spatial joins and spatial selections. For example, the Sequoia 2000 project [13] aimed at the retrieval of raster data, while the Sloan Digital Sky Survey [14] poses the need for the creation of multi-terabyte astronomy archive. The large scale systems for the analysis of remotely sensed images were specialized toward the detection of particular features like volcanoes [2], or proposed distributed and parallel data storage and query processing systems for handling of geo-scientific data retrieval queries [11]. The GeoBrowse project aims to provide infrastructure that would enable the analysis of large databases containing satellite images. Our work addresses two issues. One is the extraction of information that enables reduction of the data from multi-spectral images into a number of features. Second is the organization of the features that would allow flexible and scalable discovery of the knowledge from the databases of remotely sensed images. In this paper we present the concept of data mining system for the analysis of satellite images and preliminary results of the experiments with the collection of LANDSAT images.

## Keywords

Remote Sensing, Image Databases, Bayesian Classification, Similarity Searches.

## 1. INTRODUCTION

Satellite data is used in many different areas ranging form agriculture, forestry, and environmental studies to transportation and mining. The applications include measurements of crop and timber acreage, forecasting crop yields and forest harvest, monitoring urban growth, mapping of ice for shipping, mapping of pollution, recognition of certain rock types, and many others. The United States Geological Survey web site [15] presents other applications that use the results of the satellite data analysis.

The *GeoBrowse* project aims at providing the infrastructure required for the analysis of satellite images. Most of the systems for analyzing remotely sensed images allow simple queries based on the date of image capture and location. Such systems also allow only simple analyses of single images. When we deal with large collections of remotely sensed images, the current systems do not scale well. Therefore new algorithms and new indexing methods are needed to enable the analysis of data produced by satellite systems.

In order to facilitate the analysis of large amount of image data, we propose to extract features of images. Large images are partitioned into a number of smaller and more manageable image tiles. In addition to faster extraction of segments the partitioning allows to fetch only the relevant tiles when only retrieval of part of the image is requested. Then these image tiles are processed in order to extract feature vectors. The *GeoBrowse* architecture distinguishes between three types of feature vectors: 1) *pixel level features*, 2) *region level features*, and 3) *tile level features*. Pixel level features store spectral and textural information about each pixel of the image. For example, the fraction of the endmembers, such as concrete or water, can describe the content of the pixels. Due to the large size, pixel feature vectors are used only for the extraction of other feature vectors and can be utilized in the refinement step of the queries. Region level features describe groups of pixels. Following the segmentation process, each region is described by its boundary and a number of attributes which present information about the content of the region in terms of the endmembers and texture, shape, size, fractal scale, etc. Image tile level features present information about whole images using texture, percentages of endmembers, fractal scale and others.

There are many similarities between data mining in the collections of photographic images and data mining in the collections of sattellite images. In both cases features, such as texture, or color histograms are used in the analysis. However, in the case of the remotely sensed images a user can use additional information, such as Digital Elevation Models (DEM), or land use maps, to enhance the search capabilities and improve the quality of the classification and prediction process.

In this paper we give an overview of the GeoBrowse system and present the results of similarity searches for different types of urban areas. For the experiments we used the LANDSAT image of Western Washington State. This image contains about 500MB of raw pixel information in 6 bands (3 visible range and 3 near infrared bands). The image was corrected for atmospheric and terrain distortions, and georeferenced. In order to enable work

Proceedings of the International Workshop on Multimedia Data Mining (MDM/KDD'2000), in conjunction with ACM SIGKDD conference. Boston, USA, August 20, 2000
(O.R. Zaïane, S.J. Simoff, eds.)

with chunks of images that are feasible for the segmentation algorithm, the whole image was divided into 512 pixels × 512 pixels *image tiles*.

The remaining part of the paper is organized as follows. In Section 2 we present the architecture of the system. Section 3 describes the algorithm for the segmentation of multiband images and features that describe the regions. In Section 4 we present theresults of the similarity retrieval queries. Section 5 outlines the data mining methods for the analysis of remotely sensed images. The paper ends with conclusions and the description of future work.

## 2. ARCHITECTURE

We decided to use a database system for storage of images and their features. This way we may overcome limits related to the maximum size of files and benefit from indexing, query optimization, and partitioning features of the database. The image tiles and pixel level features are stored as BLOBs, each band in a separate column. The region and tile level features are stored in regular database tables, which can be easily accessed for further processing using GeoBrowse functions or by over 3000 function of S-PLUS software [12].

Spatial information about region level is stored in ESRI's Spatial Data Engine (SDE) together with the relevant GIS information. SDE provides open data access across local and wide area networks and the Internet using the TCP/IP protocol. It can retrieve data and perform spatial and geometric analysis with 14 topological searches, buffering, overlays and intersections, dissolve and clip, and topological data cleaning. Data stored in SDE can be also accessed from other ESRI products like ArcInfo, ArcView, and MapObjects, which provide alternative environment for the visualization of the query results.

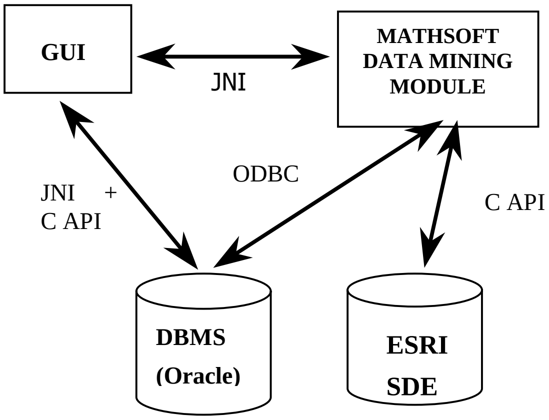

**Figure 1. Architecture of GeoBrowse**

A mining process or a similarity search is initiated by submitting a query written in a language similar to SQL-like data mining languages, such as DMQL [5] and GMQL [8]. In a query a user can specify the type of knowledge to be discovered; the set of data relevant to the mining process; and the thresholds to filter out uninteresting rules. Based on this query an SQL statement is constructed to retrieve the relevant data. If spatial conditions exist in the query the SDE is used for the processing, otherwise the data is retrieved directly from the database system. The data mining module processes the data and passes the information about the resulting tiles and regions to GUI, which in turn directly retrieves the images from the database.

Based on the classification model the data can be classified into a number of land cover classes and the resulting GIS map can be stored in the SDE for future use or a presentation by the GIS system.

## 3. SEGMENTATION AND FEATURE EXTRACTION

The segmentation process is using the function based on the algorithm presented in [7]. This function segments an input image into non-overlapping regions by minimizing an energy functional which trades off the similarity of regions against the length of their shared boundary. It starts by breaking the image into many small regions. The algorithm merges into one region the two adjoining regions that are the most alike in terms of the specified polynomial model given the length of the border between the two regions. Internally, the energy functional is evaluated using a Lagrangian parameter called lambda. Parameter is also called the scale parameter as it controls the coarseness of the segmentation where a small value of lambda corresponds to a finer segmentation with more regions and a large value corresponds to a coarse segmentation with fewer regions. Since the algorithm grows regions by merging alike regions, the value of lambda increases as the number of regions decreases. To achieve the segmentation uniformity between tiles the final value of lambda is set to be approximately the same for each image tile.

In the case of multi-band satellite images the values of the pixels are often correlated. Therefore, the Principal Component Analysis is performed based on a large sample of pixels from all tiles, all tiles are rotated to the same axes and the first three components are used for the segmentation of each image tile. After the segmentation the shape features such as eccentricity, orientation of the main axis, and invariant moments are extracted and stored in the database.

### 3.1 Texture Feature Extraction

We extract pixel level texture features based on Gabor wavelets. In the comparison study of texture based classification, Gabor features were judged to perform superior to other texture analysis methods, such as edge attribute processing methods, the circular simultaneous autoregressive model method and hidden Markov model methods [3]. In GeoBrowse for each pixel we extract eight features $a_i\big|_{i=0,7}$ using Gabor Filters with kernels rotated by $i\pi/8$. To achieve the rotation invariant features we find the values of the autocorrelation function $t_n = \sum_{i=0}^{7} |a_i| |a_{(i+n)\bmod 8}|$ [6]. To minimize the size of the pixel index, we have chosen to compute values of autocorrelation for $n = 0,2,4$. These values correspond to the 0°, 45°, and 90° difference in the orientation of Gabor kernels. Such shift should allow for distinguishing of urban road network, which usually are correlated within 90° rotation of the wavelet kernels. The extraction of other microfeatures such as frequency, orientation, is also possible [6] and we plan to perform more experiments with these features in the future.

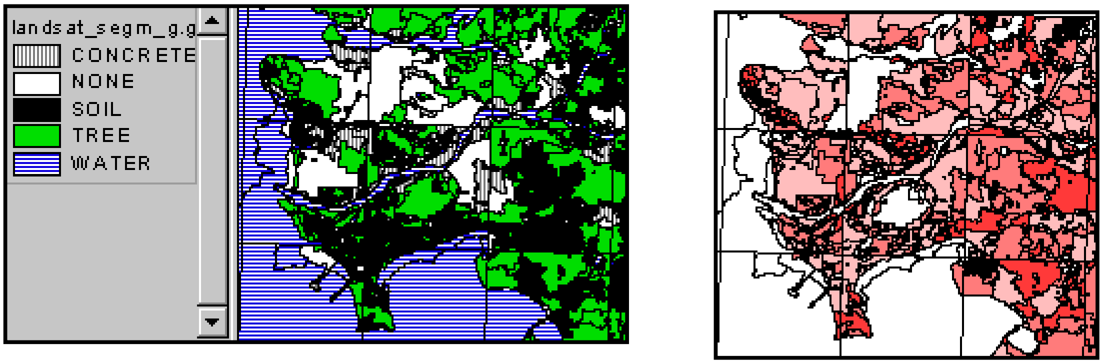

**Figure 2.**

**Spectral Mixture Analysis region features (the prevalent endmembers).**

**Percentage of pixel that belong to one of the clusters.**

## 3.2 Spectral Mixture Analysis Features

Spectral Mixture Analysis (SMA) [1, 4] enables the analysis of remotely sensed images using spectral endmembers such as concrete, water, soil, trees, etc. The pixels usually cover the area with the mixture of different endmembers. For example, in the urban areas we may find a mixture of concrete, trees, soil, grass, etc. The result of SMA represents the percentage of the contents of the endmembers within the area of the pixel. This way we can distinguish areas with different mixture of concrete, soil, water, and vegetation. The region and tile level features present the percentage of the area of a region a tile that is covered by particular endmembers. Region level texture and SMA features are presented in Figure 2.

## 4. SIMILARITY SEARCH

The GeoBrowse uses an SQL like query language that enables specification of the data mining task, features that are used in the mining process and further constraints. The system is capable of performing similarity searches based on any combination of features. A user can look for the most similar image tiles or the most similar regions based on a pattern tile or a region. GeoBrowse enables arbitrary weighting of the features. The values of the features can be adjusted to have the range [0, 1], they can be multiplied by a specific value, or they can remain the same.

In the case of region based searches we looked only for the regions with areas larger than 2000 pixels. The feature values were scaled to the range [0,1]. We compared the results of the similarity searches based on SMA features with the searches based on texture features and searches based on the combination of these two features. When only a single feature vector is used the results tend to have a high percentage of the areas, which could be classified as *false hits*. The selectivity of the SMA features seems to be quite high for urban patterns, but some rocks and crops have spectral signatures similar to the spectral signature of concrete and are classified as such. The selectivity of searches based on texture features is lower, but rotation invariance can be observed regardless of the orientation of the street networks. For example, the suburban area of New Westminster in the Greater Vancouver area is judged to be similar to East Vancouver, despite the fact that the main direction of the street network differs by about 30° for these two region. Figure 3 presents the result of the search for regions similar to downtown Seattle and Burnaby in British Columbia. Only regions in Puget Sound are shown. In the case of downtown Seattle the set of returned regions contained downtown areas of Vancouver, Burnaby, Bellingham, Bellevue, Tacoma, and Everett together with industrial areas of Renton, Tukwila and South Tacoma. Regions similar to Burnaby contain high-density residential areas with some small industrial and commercial pockets.

We compared the results of the tile similarity search with the region similarity search in the case when the tile containing the pattern region is treated as a pattern tile. In this case the returned tiles contained only about 40% of the top 20 most similar regions returned by region based similarity function. The features of the smaller regions tend to be overwhelmed by the overall features of the tile.

## 5. DATA MINING FUNCTIONALITY

In addition to the similarity search the GeoBrowse system will provide functionality for other types of the remotely sensed data analysis. This functionality will include the clustering of the data, building regression and classification models, prediction of land cover types, summarization of the data, etc.

## 5.1 Clustering

A user has an option to find clusters of image tiles based on any combination of feature vectors. Figure 4 shows the centroids (i.e., the image tiles located the most centrally in the feature space) for the four clusters. The clusters were found based on the relative content of endmembers in an image tile. In this case we may see that the image tiles that are the centroids of the discovered clusters represent mountain areas with large content of conifer trees; areas covered with deciduous trees; forested areas close to water; and urban areas close to water.

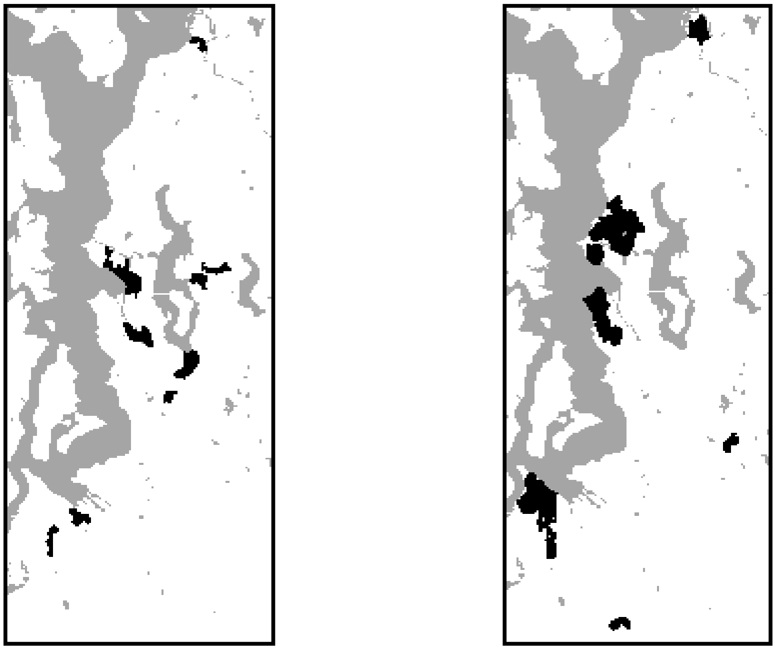

**Figure 3.**

**Regions similar to downtown Seattle.** **Regions similar to West Burnaby, BC.**

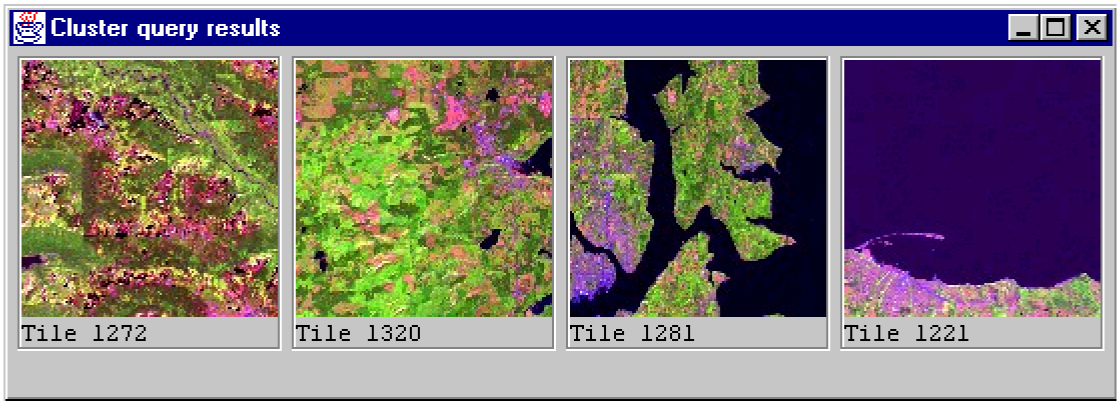

**Figure 4. The centroid tiles of the clusters.**

## 5.2 User Feedback Label Learning

In many cases it is very difficult to describe analytically the features of the objects that a user is looking for. Therefore the improvement of the description quality may play an important role in the image analysis. A method for interactive training of land cover labels using Naïve Bayesian classifiers is described in [10]. In that approach a user can interactively train Bayesian model to define a number of land cover classes, which can be based on textural or spectral properties of images. The training is done based on pixel level features, which are partitioned into a number of clusters. A user selects the pixels that belong to a new class and the pixels that do not belong there. Based on this information a model that estimates a posteriori probability of pixel's class membership is build. Using this model a user can find images with the highest probability of the defined class, or images with low or high separability of the classes. While the training is based on the pixel level features the retrieval is based on tile level features. Due to the nature of Naïve Bayesian classifier, which assumes the conditional independence of the attributes, it is possible to find out the probabilities of the pixel class assignment based on the aggregated information about all pixels in the image tile. Unfortunately the assumption of conditional independence is not always true. Therefore, Naïve Bayesian classifiers may perform well. We plan to add other classification methods, such as tree classifiers to improve user feedback label training. Because the classification process on the pixel level would be extremely expensive to compute we intend to perform experiments with the classification based on the region level features.

Building a classifier based on millions pixel features of the data would be a very time process. Instead of that we build the classifier based on region level features. In addition to spectral properties of the regions we can perform classification also based on shape properties and area of the regions, as well as auxiliary GIS information. For example, the spectral reflectance of concrete is very similar to spectral reflectance of different type of rocks. Additional information, such as Digital Elevation Models can be used to distinguish between these two types of land cover types.

## 6. FUTURE WORK AND CONCLUSIONS

We plan to perform experiments using multiple level spatial transformation methods for progressive refinement using more level than tile, region, and pixel levels. Multiscale image coding techniques, such as wavelets, can also be used for the analysis of images on multiple levels.

Such multilevel information can be combined with the auxiliary data in both vector and raster formats to enhance the data analysis capabilities of *GeoBrowse.* These auxiliary data can be used both during feature extraction process and during data mining process. We intend to do more experiments with other data mining methods such as regression, clustering and classification.

The quality of classification of land cover classes can be improved using time series of data, which can better differentiate between different types of crops due to the different times of crop growing seasons. We also plan to provide the functionality of multilevel presentation of the discovered knowledge. For example, the system should allow a user to see the generalized summary of the areas of particular crops by county, state, region, etc.

We designed, and we are in the process of implementing the *GeoBrowse* system for data mining of remotely sensed images. Three levels of feature are extracted from image tiles and used in the data mining process. In addition to simple queries based on simple properties, such as geographic location or acquisition date, a user can submit queries based on properties of images derived from feature vectors describing the images. The system also will allow for interactive training of the classification models that describe new types of objects. Scalability to the large databases is addressed through indexing of the feature vectors and by using scalable data mining algorithms in the query processing. Our region level indexing strategy enhances the data analysis and similarity search processes by allowing for the more refined classification of information derived from images.

## 7. REFERENCES

[1] Adams, J. B., M. O. Smith, and P. E. Johnson, Spectral Mixture Modeling: a New Analysis of Rock and Soil Types at Viking Lander 1. In *J. Geophys. Res*. 91:8113 – 8125, 1986.

[2] Fayyad, U. M., and P. Smyth. Image Database Exploration: Progress and Challenges. In *Proc. 1993 Knowledge Discovery in Databases Workshop*, Washington, DC, p. 14 – 27, 1993.

[3] Fountain, S. R., T. N. Tan, K. D. Baker. A Comparative Study of Rotation Invariant Classification and Retrieval of Texture Images. In *On-Line Proceedings of the Ninth British Machine Vision Conference* 1998. http://www.bmva.ac.uk/bmvc/1998/index.htm.

[4] Gillespie, A. R., M. O. Smith, J. B. Adams, and S. C. Willis, Spectral Mixture Analysis of Multispectral Thermal Infrared Images, In *Proceedings of the 2nd Thermal IR Multispectral Scanner Workshop*, JPL Publication 90-55:57 – 74, 1990.

[5] Han, J., Y. Fu, W. Wang, K. Koperski, and O. R. Zaïane. DMQL: A Data Mining Query Language for Relational Databases. *In Proc. of the Workshop on Research Issues on Data Mining and Knowledge Discovery*, Montreal, QB, pp. 27 – 34, 1996.

[6] Hayley, G. M., and B. M. Manjunath, Rotation Invariant Texture Classification using Modified Gabor Filters, In *Proc. of IEEE ICIP95*, pp. 262 – 265, 1994.

[7] Koepfler, G., C. Lopez and J. M. Morel, A Multiscale Algorithm for Image Segmentation by Variational Method, *SIAM Journal of Numerical Analysis*, vol. 31, pp. 282 – 299, 1994.

[8] Koperski, K. *A Progressive Refinement Approach to Spatial Data Mining*. Ph.D. Thesis. Simon Fraser University, 1999.

[9] Patel, J., et al. Building a Scalable Geo-Spatial DBMS: Technology, Implementation, and Evaluation. In *Proc. ACM-SIGMOD International Conference on Management of Data*, Tucson AZ, pp. 336 – 347, 1997.

[10] Schröder, M. Interactive Learning in Remote Sensing Image Databases In *IEEE Intern. Geoscience and Remote Sensing Symposium IGARSS'99*. Hamburg, 1999.

[11] Shek, E. C., R. R. Muntz, E. Mesrobian, and K. Ng. Scalable Exploratory Data Mining of Distributed GeoScientific Data. In *Proc. of The Second International Conference on Knowledge Discovery & Data Mining*, Aug. 2-4, Portland OR, pp. 32 – 37, 1996.

[12] *S-PLUS 2000 Programmer's Guide*, Data Analysis Products Division, MathSoft, Seattle, WA, 1999.

[13] Stonebraker, M., J. Frew, K. Gardels, and J. Meredith. The Sequoia 2000 Storage Benchmark. In *Proc. ACM-SIGMOD International Conference on Management of Data*, Washington, D.C., pp. 2 – 11, 1993.

[14] Szalay, A., P. Kunszt, A. Thakar, J.Gray, D. Slutz, and R. J. Brunner. Designing and Mining Multi-terabyte Astronomy Archives: The Sloan Digital Sky Survey. In *Proc. ACM-SIGMOD International Conference on Management of Data*, Dallas TX, pp. 451 – 462, 2000.

[15] U.S. Department of the Interior, U.S. Geological Survey, Landsat 7 data users and applications http://edcwww.cr.usgs.gov/l7dhf/L7MMO/L7applicatn.htm, 1999.