VisiMine: Interactive Mining in Image Databases

Krzysztof Koperski, Giovanni Marchisio, Selim Aksoy, and Carsten Tusk
Insightful Corporation
1700 Westlake Ave. N, Suite 500
Seattle, WA, 98109-3044
{krisk, giovanni, saksoy, ctusk}@insightful.com

Abstract – We describe VisiMine, a system for data mining and statistical analysis of large collections of remotely sensed images.

I Introduction

The VisiMine project provides the infrastructure and methodology required for the analysis of satellite images. In order to facilitate the analysis of large amounts of image data, we extract features of the images.

Large images are partitioned into a number of smaller, more manageable image tiles. Partitioning allows fetching of just the relevant tiles when retrieval of only part of the image is requested, and provides faster segmentation of image tiles. Individual image tiles are processed to extract the feature vectors. The VisiMine architecture distinguishes between pixel, region and tile levels of feature vectors.

Pixel level features describe spectral and textural information about each individual pixel. Polygon level features describe connected groups of pixels. Following the segmentation process, each polygon is described by its boundary and by a number of attributes that present information about the content of the region in terms of shape, size, etc. The spectral and texture properties are based on pixel features of points within the polygon. Tile level features present spectrum and texture information about whole image tiles.

All image features, together with the original images, are stored in a database system and indexed for fast retrieval. The auxiliary raster data such as Digital Elevation Models (DEM) can also be stored in the database and can be used for feature extraction and data analysis. The Oracle database system provides the means for fast information retrieval and network accessibility. Region level features can be stored in an ESRI Spatial Data Engine (SDE), and can then be displayed using ArcInfo or ArcView together with associated labels, features, or statistics. This storage functionality enables the fusion of GIS, optical, and DEM information for a variety of statistical analysis methods.

The data mining power of VisiMine includes similarity searches on tile and polygon levels, clustering of tiles, classification and regression analysis, label training using Bayesian and tree models, and connection to S-PLUS with over 3000 statistical functions. The data mining queries are specified in an SQL-like language. A user may specify the features that are used in the mining task and constraints used to select data for the mining process. The graphical query constructor enables fast query creation by non-technical users.

The user has high level of flexibility in choosing the features and images used for data analysis. The graphical user interface enables presentation of the models on high and general levels as well as drilling down into the details. The label training module enables interactive definition of models for land cover labels.

The tile level summaries of pixel features are used for fast retrieval of tiles with high/low content of features and scenes with low confidence of classification. The initial model can be refined based on the feedback supplied by a data analyst who interactively trains the model using the system output and/or additional scenes. The experts may also refine models created by other users. The VisiMine system enables construction of sophisticated statistical models using the S-PLUS system, which can access data directly from the database, or using the GUI.

II Architecture

We decided to use a database system for storage of the images and their features in order to overcome some limits on to the maximum size of files, and to benefit from indexing, query optimization, and partitioning features of the database. The image tiles and pixel level features are stored as BLOBs, with each band or feature stored in a separate column. The region and tile level features are stored in regular database tables that can be accessed easily for further processing, using VisiMine functions or other software.

Spatial information about region levels also can be 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. This information can be combined with region level features such as texture, spectral properties, or DEM features. Other ESRI products, such as, ArcInfo, ArcView, and MapObjects can access the data stored in the SDE, together with additional map data. The SDE format allows a fusion of GIS, optical, and DEM information for a variety of visualization methods and data analysis functions.

A mining process or a similarity search is initiated by submitting a query written in a data mining language similar to SQL. The query language allows the user to specify the type of knowledge to be discovered, the set of data relevant to the mining process, and the conditions that have to be satisfied by the data. Based on this query, an SQL statement is constructed to retrieve the relevant data. The data mining module processes the data and passes the information about the resulting tiles and regions to the GUI, which in turn directly retrieves the images from the database. The

capabilities of the data mining engine are enriched through the Java connection to the S-PLUS statistical data analysis engine.

The graphical user interface enables browsing and manipulation of the satellite images and associated features, creation of data mining queries, and visualization of the results of the data analysis.

The user can create new image types, and can extract image features using the VisiMine Data Manager, which is a separate program. The data manager utility also enables loading of image and DEM data from a file system to the database, browsing of meta-data, and editing of the feature descriptions. Intuitive dialogs guide the user through the specification of parameters necessary for feature extraction, the creation of data types, and the loading of data.

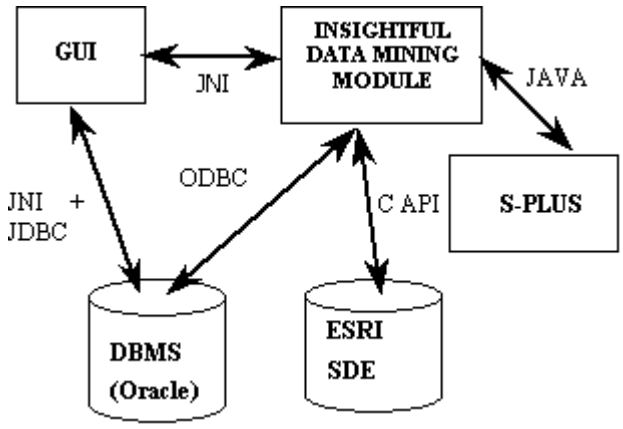

Fig. 1. The Architecture of the VisiMine System.

III Feature Extraction

The VisiMine architecture supports three levels of features: pixel, region, and tile level features. The feature extraction process starts with the analysis of spectral and textural properties at the pixel level. The numerical pixel data can be clustered in order to find a small number of classes. At the same time, tile level features may be extracted, thereby creating histograms of the pixel classes for each tile.

The extraction of region level features starts with a segmentation algorithm. The geometrical properties of regions, such as image moments, are extracted. Based of the pixel features, the system computes statistical properties of regions such as histograms, max, min, mean, and standard deviation features for each region. Additional features are extracted using raster information, such as digital elevation maps. These features can also be created at all three levels.

The VisiMine system supports extraction of the following features:

- Texture features using Gabor wavelets, Haralick's co-ocurrence, and Laws texture.
- Clustering (spectral, textural, and others) using: CLARA [2] (medoid algorithm), RHSEG [4] (hierarchical algorithm), and k-means [2].
- Spectral Mixture Analysis features.

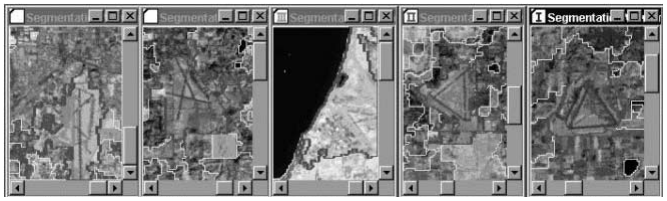

Fig. 2. Region Similarity Search for Airports.

- Segmentation and shape descriptors of the regions.
- Spatial relationships between regions.
- Histograms, max, min, mean, and standard deviation of pixel features for each region and tile.

IV Data Mining Functionality

VisiMine uses an SQL-like query language that enables specification of the data mining task, features to be used in the mining process, and any additional 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 for the most similar regions based on a pattern tile or a pattern region. VisiMine allows 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 unchanged.

Fig. 2 presents the result of a search for regions similar to McCord Air Force Base. Among the top 4 most similar regions are Olympia Municipal Airport, Whidbey Island Naval Air Station, Arlington Airport, and Abbotsford Airport.

We compared the results of the tile similarity search with the region similarity search for the case when the tile containing the pattern region is treated as the pattern tile. In this case, the returned tiles contain only a small fraction of the most similar regions that were returned by region based similarity function. The features of the smaller regions tend to be overwhelmed by the overall features of the tile.

In addition to the similarity search, the VisiMine system provides functionality for other types of analyses of remotely sensed data. This functionality includes data clustering, building regression and classification models, prediction of land cover types, summarizing data, searching using visual grammar and interactive label training using Bayesian and tree models.

Interactive label training methods enable searches for features that are very difficult to describe analytically. In VisiMine we use a method for training of land cover labels that employs naïve Bayesian classifiers described in [3]. In this approach a user can interactively train a 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 partitioned into a number of clusters. First, beginning with a single image tile, a user provides positive examples by selecting regions with pixels that belong to the same class, and the system then builds a model for use in identification of additional regions belonging to that class. He/she also has to provide negative

examples, by selecting regions with pixels that belong to other classes. Based on this information, a model that estimates the posteriori probability of a pixel's class membership is built. The probability of pixels for the selected tile is shown on the screen. When a user judges the model to be appropriate for pixels in the currently selected tile, he/she poses a query that finds either images with a high probability of belonging to the defined class, or images with a low probability (high separability from the class). Using these images, the user may choose to continue training based on other image tiles until the model is sufficiently refined.

While the training is based on pixel level features, the retrieval is based on tile level features. Due to the nature of the naïve Bayesian classifier, which assumes conditional independence of the attributes, it is possible to find the average probabilities of the pixel class assignments in a tile based on the aggregated information about all pixels in the image tile. Despite the fact that the assumption of conditional independence is not always true, naïve Bayesian classifiers perform well in practice.

Another method for label training implemented in VisiMine is based on decision tree models. Because the classification process on the pixel level would be extremely expensive to compute, the decision tree models are based on region level features. In addition to spectral properties of regions we can perform classification based on shape properties and areas of regions, as well as on auxiliary GIS information. In a way similar to Bayesian label training, a user provides the system with positive examples by pointing to regions that belong to the trained class, and negative examples by pointing to other regions.

V S-PLUS CONNECTIVITY

Insightful's S-PLUS is an interactive computing environment for graphics, data analysis, statistics, and mathematical computing. It contains a superset of the S object-oriented language and system originally developed at AT&T Bell Laboratories, and it provides an environment for high-interaction graphical analysis of multivariate data, modern statistical methods (e.g., robust and non-parametric methods), data clustering and classification, and mathematical computing. In total, S-PLUS contains over 3000 functions for scientific data analysis. VisiMine data can be accessed from within S-PLUS by using Java connectivity for images and ODBC connectivity for image and region data. In addition, VisiMine has the S-PLUS command tool, which provides for easy transfer of images to S-PLUS, and for data processing using the S-PLUS language. The S-PLUS images can be returned to VisiMine and displayed there.

The VisiMine can also display S-PLUS graphics, which are created using a command line interface and shown within S-PLUS plot window. Fig. 3 shows a graph of the brightness and greenness indices for an agricultural part of British Columbia. The Kauth-Thomas tasseled cap for vegetation in the brightness-greenness plane [1] can be recognized, together with clusters of water and bare soil pixels.

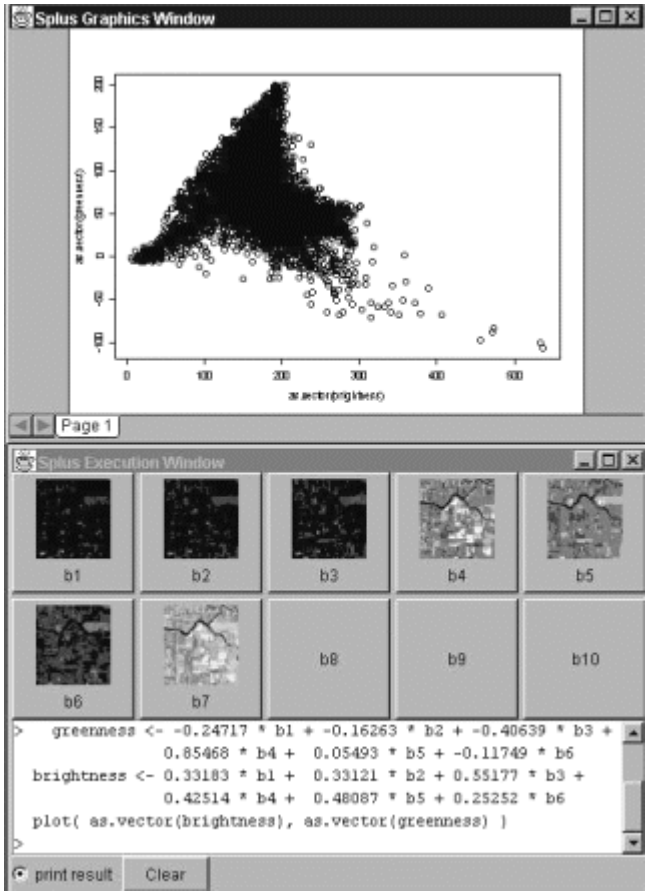

Fig. 3. S-PLUS Interface and Graphics.

The combination of S-PLUS and VisiMine features creates a unique environment for interactive exploration and analysis of remotely sensed data. The rich statistical functionality of S-PLUS, together with the VisiMine user interface and the scalability of its data mining engine, allows for easy and powerful customization of the data analysis process.

ACKNOWLEDGEMENTS

Funding for the prototype comes from NASA SBIR Phase II contract NAS5-98053 and by NRA2-37143 contract.

REFERENCES

[1] Jensen, J. R. Introductory Digital Image Processing: a Remote Sensing Perspective. Prentice-Hall, 1996.

[2] Kaufman, L. and P. J. Rousseeuw. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York, 1990.

[3] Schröder, M., H. Rehrauer, K. Seidel, and M. Datcu. Interactive Learning and Probabilistic Retrieval in Remote Sensing Image Archives. In IEEE Trans. on Geoscience and Remote Sensing, Sep. 2000, Vol. 38, No. 5 pp. 2288 - 2298.

[4] J. C. Tilton, "A recursive PVM implementation of an image segmentation algorithm with performance comparisons between the HIVE and the Cray T3E." Proc. of the 7th Symposium on the Frontiers of Massively Parallel Computation, Annapolis, MD pp. 146-153, Feb. 21-25, 1999.