Querying Remote Sensing and GIS Repositories with Spatial Association Rules

Giovanni B. Marchisio, Krzysztof Koperski and Michael Sanella
Data Analysis Products Division of Mathsoft, Inc.
1700 Westlake Ave N, Suite 500, Seattle, Washington, 98109-3044 USA
Tel: (206) 283-8802, Fax: (206) 283-8691
email: giovanni@splus.mathsoft.com, krisk@splus.mathsoft.com, sannella@splus.mathsoft.com

We describe the fusion of multispectral image and GIS data mining functions in a module that implements spatial association rules. Spatial association rules may represent: 1) topological relationships between spatial objects; 2) spatial orientation or ordering; and 3) distance information. Our approach relies on a fast and unique multichannel segmentation algorithm that combines region-based information with edge-based information in a variational framework. The indexing strategy distinguishes between three levels of features: 1) pixel level, 2) region level, and 3) scene level features. We use pixel level information for the extraction of higher level features, and in the process of query refinement. Region level features describe groups of contiguous pixels. Following segmentation, we describe each region with a boundary and a number of attributes, like spectral endmember types and percentages, textural classes, shape, size, fractal scale, etc. We quantify shape, orientation and other geometric properties of the segments by computing 32 moments. Scene level features describe global properties of whole scenes, and the spatial relationships of the largest regions in them. The second level of features can optionally support definition of semantic labels from multiple attributes, which include GIS attributes, if present. At a higher level, we can attempt to construct semantic labels from a tree of topological relationships. We store images and features in a database to enable fast access, data integrity and easy interfacing with other applications.

STORAGE AND DATA FUSION

Our data mining and information retrieval module [1] fuses information derived from remotely sensed imagery with GIS information. Initially, multispectral images are stored as BLOBs in a relational database. An automatic indexing process attempts to extract a variety of interesting statistical descriptors from the images and stores them as feature indices in the database. For each tile we extract pixel level, region level and scene level features. At an increasing level of information abstraction, the storage representations that correspond to each level are raster, vector formats and statistical data frames. We achieve data fusion at the second and third level of feature representations. We employ a fast and robust algorithm for automatic region segmentation based on multiscale or PDE methods [2]. Following [3], we minimize a simplified Mumford and Shah functional. CPU times for the segmentation of a 512x152 Landsat TM image are twenty seconds on a Sun Ultra 60 workstation across six spectral bands. This enables us to segment large databases of images covering large areas in a relatively short time. As an example, in Fig.1 we show the multispectral region segmentation for a 512x512 image tile of Vancouver, Canada.

Fig. 1: a multispectra image of Vancouver, Canada (left) and the results of the fast region segmentation used in the automated indexing process (right).

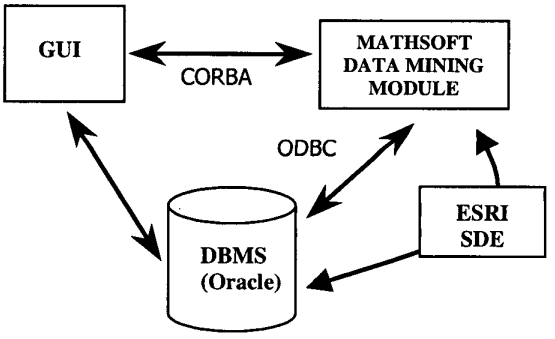

Fig. 2: System Architecture

The spatial descriptions of the regions created by the segmentation process are stored in the ESRI's Spatial Data Engine (SDE) along with other vector GIS information. Our open architecture, which is shown in Fig.2, allows uniform storage and access of image derived information from a variety of other application modules familiar to the remote sensing community. We have designed multispectral data structures which allow the simultaneous representation of classified image, region image, and edge images in a single data structure. Image derived features can be used in conjunction with GIS data to build statistical models that explain the relationships between data objects, or to perform information retrieval experiments. SDE supports spatial

0-7803-6359-0/00/$10.00 © 2000 IEEE

operations, such as the creation of spatial buffers for proximity operations and spatial joins that can be used to create predicates describing spatial relationships between objects in the database.

FEATURES

Each region in the database contains statistical descriptors that characterize image derived information such as the relative abundances of spectral reflectances, textural classes, textural orientation, size, fractal scale and shape. We perform localized Spectral Mixture Analysis (SMA) [4] and store relative endmember fractions for each region. The preferred method of textural characterization is based on Gabor wavelets. These seem to outperform other methods for texture analysis, such as edge attribute processing, the circular simultaneous autoregressive model and hidden Markov model methods [5]. For each pixel we extract eight features a_i $(i=0,7)$ with Gabor kernels rotated by $i\pi/8$. Following [6], we compute rotational invariant features by taking the values of the autocorrelation function $t_j = \sum_{i=0}^{7} |a_i| \times |a_{\mathrm{mod}(i+j,8)}|$. To minimize the size of the index, we have chosen to compute values of the autocorrelation for $j=0,2,4$. These values correspond to the 0°, 45°, and 90° difference in the orientation of Gabor kernels. This selection is sufficient for detecting urban road networks. Our approach also allows for the extraction of other microfeatures such as frequency and orientation. We show an example of region level texture features in Fig.3 for an area of Greater Vancouver, Canada covered by 9 Landsat TM tiles. In this example, we cluster the values of autocorrelation for each pixel into 12 groups. The region and scene level feature indices store the percentage of pixels within each particular region and scene that belong to each of the clusters. We present SMA features corresponding to the same area of Fig.4. We use four endmember reflectances and compute their relative abundance in each segment. The map labels each region by the dominant endmember. If there is no dominant endmember, as it is the case in some suburban areas, the label is "none".

We quantify the shape of each region by computing and storing 32 moments. Our indexing methodology enables the fusion of image derived information with auxiliary GIS information, such as land use, zoning, hydrographic or transportation network overlays, and demographic information. The data mining functions enable joint statistical analysis and similarity searches in these hybrid data repositories. For instance, a user can perform queries by image content (QBIC) with the constraint that the retrieved region(s) must be close to a major highway.

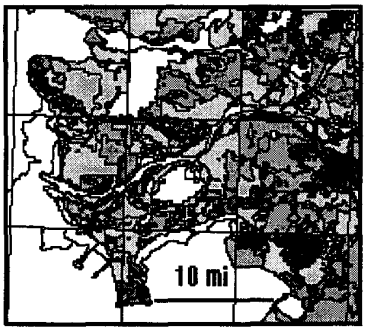

Fig. 3: Region Level Texture Features.

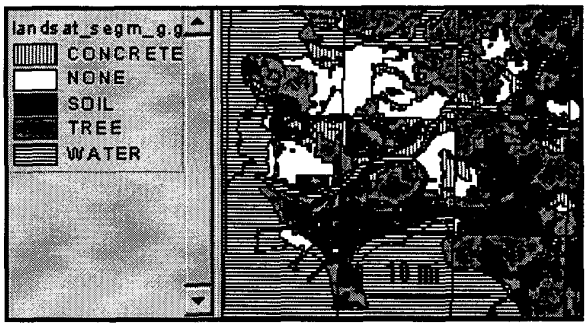

Fig. 4: Region Level SMA Features.

SPATIAL ASSOCIATION RULES

A spatial association rule is of the form $X \bullet Y$ (c%), where X and Y are sets of spatial or non-spatial predicates and c% is the confidence of the rule [7]. An example of a spatial association rule is: *prevalent_endmember(x, concrete)* $\wedge$ *texture_class(x,C1)* • *close_to(x, coastline)* (60%). This rule states that 60% of regions where concrete is the prevalent endmember and that texture features belong to class C1 are close to the coastline. There are various kinds of spatial predicates that could constitute a spatial association rule. Examples include topological relations such as *intersects*, *overlap*, and *disjoint*, spatial orientations such as *left_of* and *west_of*, and distance information such as *close_to* or *far_away*. Spatial association rules in databases of remotely sensed images could be discovered at the pixel, region, or scene levels. The process of finding spatial association rules at the pixel level is computationally intensive. On other hand, association rules at the scene level may be too general. We concentrate therefore on finding spatial association rules at the region level. We initiate the mining process with a query that specifies the objects, the attributes, and the predicates to be used in the rules. *Minimum support* and *minimum confidence* thresholds filter out associations that describe small percentage of objects and rules with low confidence. To minimize the number of costly spatial computations performed, the algorithm first uses various approximations, like Minimum Bounding Rectangles (MBRs), to find coarse predicates, then it applies finer but more expensive spatial

computations only to those patterns having large support at the approximation level [7].

INFORMATION RETRIEVAL EXPERIMENTS

Our information retrieval module employs an SQL like query language that allows the user to specify the data mining task, feature(s) to be used in the mining process and input additional constraints. The system can perform similarity searches based on any combination of features. We can perform QBIC searches at the scene or region level, weight features differently and introduce thresholds in the search. For instance, we may be interested in retrieving only areas that are larger than 2000 pixels.

In two QBIC experiments on a database covering 3,000 square miles in Western Washington state and British Columbia we compare results based on single and multiple features. When only a single feature vector (SMA) is used the results tend to have a high percentage of areas that could be classified as *false hits*. The selectivity of searches based on SMA features appears to be high for urban areas. However, rock outcrops are often mistaken with concrete due to the similarity in spectral signatures. The selectivity of searches based on our texture features is lower, but rotation invariance can be observed, as urban regions are judged to be similar, regardless of the orientation of 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. Fig. 5 presents the result of a search for regions similar to downtown Seattle and Burnaby in British Columbia. These searches are at the region level and are based on both texture and SMA features. When the query token is an image sample from downtown Seattle, the set of returned regions include many downtown sections of Vancouver, Burnaby, Bellingham, Bellevue, Tacoma, Everett together with industrial areas of Renton, Tukwila and South Tacoma. On the other hand, regions similar to Burnaby include high density residential areas with some small industrial pockets. We also compared results of similarity searches at the region level with searches at the scene level. In this case the returned scenes contain only about 40% of the top 20 regions returned by a region level similarity search. The features of the smaller regions tend to be overwhelmed by the overall features of the scene.

CONCLUSIONS

Our module for spatial data mining allows the joint statistical analysis of data derived from remotely sensed imagery and GIS information. A key innovation is the support for "content based" searches based on properties of multispectral images. Image properties are derived on three levels: pixel, region and image tile. Our region level indexing strategy enhances the data analysis and similarity search processes by allowing for the more refined classification of image derived information.

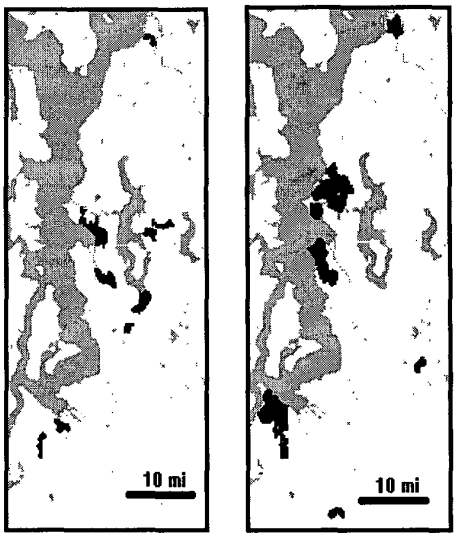

Fig.5: Regions Similar to Downtown Seattle, WA (left) and to West Burnaby, BC (right)

The system also allows for interactive training of classification models that describe new object types.

ACKNOWLEDGEMENTS

Funding for the prototype comes from NASA SBIR Phase II contract NAS5-98053. Data was provided by PRISM.

REFERENCES

[1] G. B. Marchisio and Alan Q. Li, "Intelligent System Technologies for Remote Sensing Repositories", in Information Processing in Remote Sensing, World Scientific Publishing, 1999.

[2] G. B. Marchisio and J. Cornelison, Content-based Search and Clustering of Remote Sensing Imagery Proceedings of IEEE IGARSS'99, June 1999

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

[4] J.B. Adams, 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.

[5] S. R. Fountain, 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

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

[7] K. Koperski and J. Han. Discovery of Spatial Association Rules in Geographic Information Databases. In *Advances in Spatial Databases, Proc. of 4th Symp. SSD'95*, Springer-Verlag, Berlin, 1995, p. 47-66.