Automated Feature Selection through Relevance Feedback

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

***Abstract*-The VisiMine [2] 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 and selection of the features that would allow flexible and scalable discovery of the knowledge from the databases of remotely sensed images. The VisiMine architecture distinguishes between three types of feature vectors: pixel, region and tile.**

One of the challenges in information retrieval is the proper choice of the set of features that are the best suited for a data mining task. The VisiMine system enables extraction of a large number of features that describe textural and spectral properties of satellite information. In addition to the analysis of image information, the system can perform data fusion of image properties with auxiliary data such as DEM. In [1], we presented the results of the information retrieval experiments with the Hierarchical Segmentation (HSEG) algorithm that produces a hierarchical set of image segmentations. The results presented in last year's paper showed that the use of HSEG features improves the precision and recall of similarity searches. However, for different types of land cover, different combinations of HSEG segmentation levels and textural features provided the best results. Image analysis applications often require different levels of image segmentation detail as well as the use of different mixes of spectral, textural and shape features combined together with auxiliary information. Furthermore, a particular application may require different features and different levels of image segmentation detail depending on how the image objects are being analyzed. Thus, an automatic selection of feature sets would be very useful for satellite image analysis.

In this paper, we present algorithms that allow for automatic selection of features for region and tile similarity searches. The relevance feedback technique allows for selective choices to be made in the region(s) of interest for which a good subset of features may be found in real time. The preliminary results of the experiments with LANDSAT data show improvements in both precision and recall over previously used methods.

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 from 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 also 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, and the like. 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. 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, label training using Bayesian and tree models, and connecting to S-PLUS which has 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 identify any and all constraints used to downselect data for the mining process. The graphical query constructor enables fast query creation by non-technical users.

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

The classifiers implemented in the VisiMine system include Naïve Bayes, decision trees, and minimum distance classifiers. 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

scenarios. The experts may also refine models created by other users. Users can trace rules to pixels and then pixels to rules and, ergo, change the parameters. The VisiMine system enables construction of sophisticated statistical models by using either the S-PLUS system, which can access data directly from the database, or via the GUI.

II 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 on the pixel features, the system computes statistical properties of regions such as histograms, maximum, minimum, 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 and Haralick's co-occurrence.
- Clustering (spectral, textural, and others) using: CLARA (medoid algorithm), RHSEG (hierarchical algorithm), and k-means.
- Spectral Mixture Analysis features.
- Segmentation and shape descriptions of the regions.
- Spatial relationships between regions.
- Histograms, maximum, minimum, mean, and standard deviation of pixel features for each region and tile.

III BASIC RELEVANCE FEEDBACK

With the large number of extracted features and their combinations, it is difficult for a user to choose the best combination of features. Relevance feedback techniques enable automatic weighting of the features, which in turn, may enable automatic selection of the best features for the retrieval of relevant image tiles and regions. The basic relevance feedback implementation in VisiMine uses an approach based on query vector shifting and feature selection by weighting. The query vector shifting is done using a modified version of the Rocchio algorithm [3].

Let F denote the feature space of our collection, $F = \{\vec{f}_1,...,\vec{f}_n\}$, where $\vec{f}_1 = [x_{i1},...,x_{idi}] \in \Re^{di}$ are features that have been derived from the original image data.

For each subset $S = \{\vec{f}_{\Pi 1},...,\vec{f}_{\Pi k}\} \subseteq F$, we define a *combined feature vector* $\vec{f}^S$ as:

$$\vec{f}^S = [x_{\pi 11},...,x_{\pi 1d\pi 1},...,x_{\pi kd\pi k}] \in \Re^{\sum_{i=1}^{k} d_{\pi k}}$$

A basic similarity search computes the total feature vector $\vec{q}$ for a given subset $Q \subseteq F$ for the pattern query tile or region and then compares this *query vector* to the combined feature vectors over Q for the rest of the image database. Based on the result set of this initial nearest neighbor search, the user chooses positive and negative examples and reiterates the 'search process'.

Let $R = \{\vec{r}_1,...,\vec{r}_n\} \subseteq Q$ be the result set of the search, $P = \{\vec{p}_1,...,\vec{p}_n\} \subseteq R$ be the set of positive examples and $N = \{\vec{n}_1,...,\vec{n}_n\} \subseteq R$ be the set of negative examples, and $P \cap N = nil$.

A. Query Vector Shifting

We compute the new shifted query vector based on the positive and negative sample vectors as follows:

$$\vec{q}' = \vec{q} + \beta(\frac{\sum_{i=1}^{r} \vec{p}_i}{r} - \vec{q}) - \gamma(\frac{\sum_{i=1}^{s} \vec{n}_i}{s} - \vec{q})$$

The parameters β and γ control the influence of negative and positive feedback examples. (*Basically, we shift the query vector based on a 'center of mass' computation over all positive and negative samples*).

B. Feature Selection

For the simple relevance feedback approach, feature selection is done by weighting. We compute the variance for each element of the total feature vectors in the positive $\vec{\sigma}^p = [\vec{\sigma}_1^p,...,\vec{\sigma}_n^p]$ and negative $\vec{\sigma}^n = [\vec{\sigma}_1^n,...,\vec{\sigma}_n^n]$ feedback sets where $\sigma_i^s = \frac{1}{s} \cdot \sum_{i=1}^{s} (n_i^j - \overline{n}^j)^2$.

The basic idea is that feature elements with a low variance among the positive examples are more likely to identify similar image tiles or regions than those with a large variance. Therefore, we initially define the weights as follows:

$$\text{if } \sigma_i^p > 0.01 \text{, then } w_i = \frac{1}{\ln(1+\sigma_i^p)} \text{, else } w_i = \frac{1}{\ln 1.01}$$

In order to introduce the negative feedback results into the weighting and in order to include a better feature selection/reduction, we compute the distance vector between the means of the positive and negative feedback examples as:

$$\vec{d}_{mean} = \left| \frac{\sum_{i=1}^{r} \vec{p}_i}{r} - \frac{\sum_{i=1}^{s} \vec{n}_i}{s} \right|$$

If the variance and the distance of the means for a certain feature is almost equidistant to the positive and negative example sets, we discard that feature by setting its weight to zero since it is not able to distinguish between positive and negative examples.

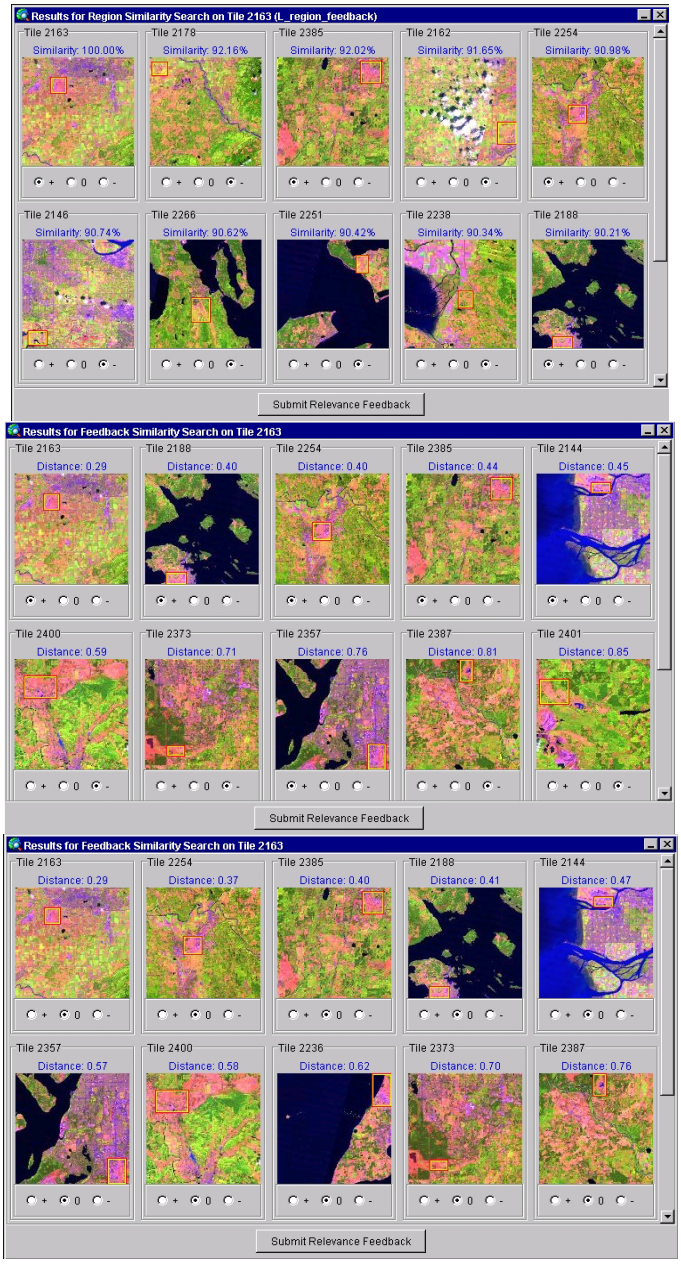

Figure 1. Relevance Feedback Search for airport.

If $d_{mean_i} < 0.01 \wedge (\sigma_i^p - \sigma_i^n) < 0.01$, then $w_i' = 0$, else $w_i' = w_i$.

The distance between modified query vectors is the generalized Euclidean distance. For each image I_k within the database with a corresponding total feature vector $v_k \in Q$, we compute $d_k = \left|(\vec{q}' - \vec{v}_k)^T \cdot diag(\vec{w}') \cdot (\vec{q}' - \vec{v}_k)\right|$.

IV Observations

First evaluation results show that the influence of negative feedback examples in the query vector shifting (controlled by parameter γ) leads to unbalanced results. It seems to push the query vector away from the actual positive examples into 'unknown space', which, in turn, leads to unwanted results. Queries with $\gamma = 0$ provide more consistent results.

The feature selection introduced with the proposed weighting scheme seems to work well. On average, we find that 10-40 percent of the initially used features are obsolete for the query, meaning that they are unable to distinguish between good and bad matches. In extreme circumstances, we have observed up to 90 percent in feature reductions. In the presentation, we will demonstrate detailed results using tables and graphs. Unfortunately, space requirements do not allow inclusion of a complete set of detailed results in this paper.

The current system works with a static set of features that are chosen when the query is composed. Future versions will begin with this initial feature set and will be expanded after the first round of users provide feedback on the full set of features that are available in the database. Thus, an optimal subset of features for the query will be chosen automatically by the VisiMine System and, in so doing, this burden will be lifted from the user. The user is very often not capable of choosing the best set of low-level features for his / her particular query. In the presentation, we will show the results of this implementation.

Acknowledgements

Funding for the prototype comes from the NASA NRA2-37143 contract.

References

[1] J. Tilton, G. Marchisio, K. Koperski, and M. Datcu. "Image Information Mining Utilizing Hierarchical Segmentation". In *Proc. IEEE International Geoscience and Remote Sensing Symposium IGARSS-02*, Toronto, Canada, July 2002.

[2] K. Koperski, G. Marchisio, C. Tusk, and S. Aksoy. "Interactive models for semantic labeling of satellite images". In *Proc. International Symposium on Optical Science and Technology SPIE's 47th Annual Meeting*, Seattle, July 2002.

[3] J.J. Rocchio. "The SMART Retrieval System". In *Relevance Feedback in Information Retrieval*, pp. 313-323, Prentice Hall, Englewood Cliffs, 1971.