

Very Large Scale Multidimensional Data Management and Retrieval for USGS and NIMA Imagery *

Aidong Zhang and Wei Wang
Department of Computer Science and Engineering
State University of New York at Buffalo
Buffalo, NY 14260 USA

David M. Mark
Department of Geography
State University of New York at Buffalo
Buffalo, NY 14260 USA

1 Introduction

Content-based image retrieval using low-level features such as color, texture and shape has been well studied. Various image querying systems have been built based on the low-level features for general or specific image retrieval tasks. The application of these approaches in geographic images have been explored, e.g. [1]. However, retrieving images based on low-level features may not be satisfactory. With the enormous growth of GIS images, it is an urgent need to build image retrieval systems which support both low-level (feature-based) and high-level (semantics-based) querying and browsing of images.

In this proposed research, our focus is to apply various approaches to automatically extract various levels of features including low-level texture features, global statistical features, and high-level semantic keywords from GIS images. These features are then clustered and indexed to extract the metadata. Housed in a metaserver, the metadata can be used to support effective and efficient querying and analysis in the distributed and heterogeneous environment of GIS images.

Our approach includes two major components. The first component is the remote GIS image databases and respective image retrieving servers. The second component is the metaserver, including the metadata database, the metasearch agent, and the query manager. The components are illustrated in Figure 1.

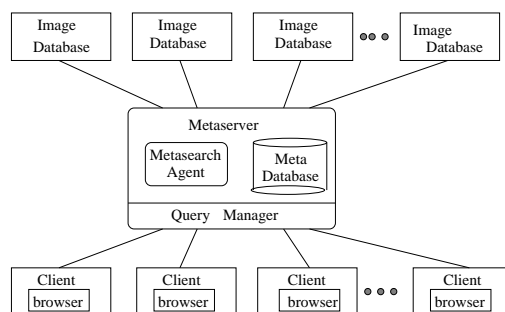


Figure 1: System architecture

*This research is supported by the National Science Foundation Digital Government Grant EIA-9983430.

The whole system works in this way. The *Query manager* extracts the image features and semantic keywords from the user queries and for suitable matching with the metadata housed in the *Metadatabase*. The *Metasearch agent* then produces a ranked list of the remote databases relevant to the queries, and guides the queries to the selected databases to retrieve images.

2 Feature and Metadata Extraction

Feature Extraction

To conduct various retrieval and analysis of GIS images, we need to extract different levels of features to describe the content of the GIS images.

Low-level texture features. Texture features have been demonstrated to be successful in representing GIS images. We use the Wavelet-based feature to represent the texture. We develop multi-resolution representation methods that are most appropriate for GIS images based on wavelet transforms.

Global statistical features. We use the Keyblock-based feature [1] to represent the global statistical features. Such statistical features will be highly useful for environmental scientists to retrieve various global information on the content of GIS images for analysis.

Semantic keywords. We use the monotonic tree [3] approach to generate the semantic keywords for GIS images. This approach has been successfully applied to extract the scenery semantic keywords.

Metadata Extraction

By clustering the feature space of GIS images, we extract the metadata which are densely populated regions in the feature space. We do so by identifying the arbitrary shaped clusters in the feature space of GIS images at different degrees of accuracy.

We apply *WaveCluster* [2], a grid-based approach to clustering multi-dimensional data, to generate the clusters of the feature space. The clustered image feature space can be represented by a set of centroids of the clusters, denoted *templates*. The clusters can be further classified into subclusters, which can then be represented by their centroids. The benefit of this approach is that a hierarchical index can be built on the templates.

3 Efficient Image Querying and Retrieval

Based on the metadata, the remote GIS databases can be ranked by the metasearch agent with regard to a particular query. To rank the databases, the similarity between the query and the metadata will be calculated to determine the most relevant metadata for the query. The metaserver will then pose the query to the relevant databases selected by users, which will then retrieve the relevant images and return them back to users.

References

- [1] L. Zhu, A. Rao and A. Zhang. Keyblock: An approach for content-based geographic image retrieval. In *Proceedings of First International Conference on Geographic Information Science (GIScience 2000)*, pages 286–287, Savannah, Georgia, USA, October 28-31 2000.
- [2] Sheikholeslami G. and Chatterjee S. and Zhang A. *WaveCluster: A Wavelet-Based Clustering Approach for Multidimensional Data in Very Large Databases*. *The VLDB Journal*, 8(4):289–304, February 2000.
- [3] Y. Song and A. Zhang. Monotonic tree. In *Proc. 10th Int. Conf. on Discrete Geometry for Computer Imagery, Bordeaux, France, 2002*.