HyperMINE – An Earth Observation Spatio-Temporal Data Mining System

Alexandru-Cosmin Grivei, and Mihai Datcu, *Fellow, IEEE*

***Abstract* — The increasing number of satellite missions for Earth exploration provides huge amounts of data. These data are of wide diversity regarding the images characteristics, thus new techniques and tools need to be developed to accommodate the extraction of meaningful information. This paper presents An Earth Observation Spatio-Temporal Data Mining System (HyperMINE), build on a modular multilayer architecture that allows effective processing of various sources of data. The designed system integrates fast and complex query methods in order to generate SITS (Satellite Image Time Series) regarded as a data hypercube. The multidimensional data, considering geographical space, time, band, and satellite/sensor are used as input for information mining algorithms.**

***Keywords* — Data mining, geospatial, SITS.**

I. Introduction

THE large quantity of data acquired by current and past satellite missions leads to important limitations when considering the problem of efficiently storing and, indexing those data, followed by the retrieval of particular information useful in different applications.

Developing a system which will accomplish these tasks in order to facilitate the use of data in scientific research or specific applications is not a trivial task. This is due to the genuine volume of information that needs to be extracted from the increasing quantity of data. In addition, by each new satellite mission the data gets more diverse, both in terms of quality and quantity of metadata accompanying each acquisition.

A system should be adaptable in order to accept new types and sources of data as input. This is why a modular architecture is a very good solution for fast development of functions, as stated in [1] or a multilayered approach, as suggested in [2] or in [3], by defining a specific functionality and logic for each layer, the system is well organized and it is easy to add new capabilities. HyperMINE takes advantage of both approaches giving it flexibility towards the addition of new functionalities and makes possible to develop modules that can permit it to accept new data sources and formats for ingestion in the system.

Several other existing systems or proposed solutions for storing the huge amount of data and for the extraction of relevant information will be presented in chapter II, while in chapter III the architecture for our proposed solution, HyperMINE will be detailed with all of its composing modules and layers.

Chapter IV will present examples of HyperMINE functionality at the time of writing the article, as it is a developing system and new modules are still being added. Finally in chapter V conclusions will be drawn regarding the proposed system and future development approaches will be presented.

II. Concepts and Systems for Image Mining: An Overview

In the past, several systems have been developed in order to serve specific applications, for different domains. Some of them use just the metadata for the indexing process. There are also more advanced approaches in which the image features are used to represent the data in databases. The best performances are obtained employing the latter approach, but usually their developers aim at optimizing the system for a certain set of images and application scenario. Thus the algorithms used may not obtain the same results if the system is used for another scenario.

As the flexibility of a system increases, the complexity rises in order to cope with the possible scenarios; it shall guarantee suitable results for any other type of user. Two such systems, which aim to obtain the best results in a reduced amount of time, are SemQuery (Semantic Clustering and Querying on Heterogeneous Features for Visual Data)[4] and GeoIRIS (Geospatial Information Retrieval and Indexing System)[5]. They try to achieve this by proposing different indexing and image retrieval techniques.

SemQuery is a data indexing system based on features such as color histograms and textures extracted from images. It supports CBIR (Content Based Image Retrieval), accepting a sample image as a query and retrieves similar images and displays them in a ranked order related to the degree of similarity. The system was

The work has been funded by the Sectoral Operational Programme Human Resources Development 2007-2013 of the Ministry of European Funds through the Financial Agreement POSDRU/187/1.5/S/155536.

Alexandru-Cosmin Grivei is with the Faculty of Telecomunications and Information Tehnology, University Politehnica of Bucharest, Bul. Iuliu Maniu 1-3, Bucharest, Romania (e-mail: alex.grivei@ymail.com).

Mihai Datcu is now with the …

evaluated using 29,400 texture and color vectors from images. The database was divided in five semantic categories: cloud, floral, leaves, mountain and water which were used in the database R-tree indexing process. For training the system, and finding the templates, 10% of the data was used. For the extraction of features, the wavelet transform was used to extract texture features, and for the color feature, the color histograms were used. The system also developed a merging and ranking method for the results.

GeoIRIS is a system developed for data modeling, indexing and extraction from large image databases. It incorporates an automatic feature extraction which retrieves a set of characteristics from the images provided for queries or for database insertion. GeoIRIS has an interface which presents the query results in an ordered manner, according to the relevance. The system uses two other sources which are merged together with the information from the GeoIRIS database to present a complex and complete data format.

The architecture of GeoIRIS includes several modules with well-defined roles. These are: the feature extraction module, the indexing structures, the semantic framework, the GeoName server, the fusion and ranking module and the retrieval and visualization system. The main features extracted from the tiles are histograms, gray scale, RGB, near-infrared data, anthropogenic features, object-based features. The indexing is done using EBS (Entropy Balanced Statistical) k-D Tree for continuous features and the EBB (Entropy Balanced Bitmap) Tree for the binary features. The system supports several query methods like query by example, object queries, hybrid queries and semantic queries.

A similar image retrieval system for Terasar-X Radar Satellite imagery was proposed in [6], using a modular and three-layer architecture. They used both metadata, and features characterizing the images in order to build up the queries needed in the data search.

Other papers propose solutions for different parts of a data mining system. For example in [7] a solution was given for an efficient indexing method for SITS, which is also the main form of data used in our proposed system. In [8] a fast visualization and data extraction solution for large data sets, VisReduce was proposed as a novel and scalable approach for text based data, showing impressive performances for the almost 150.000.000 records used for the tests.

Each of the presented systems gives solutions more or less comprehensive for the challenging problem of big data manipulation. Some propose solutions for efficiently indexing data features, such as [4] and [7], or simple text data as in [8]. Others try to give a more complex solution with an increased number of functionalities such as in [5] and [6] but do not really enter the big data realm as the test data set is not big enough.

In the next chapter the architecture of the proposed system will be presented, a system which aims to surpass some of the issues present in the above mentioned systems and methods.

III. Architecture of HyperMINE – Spatio-temporal Data Mining System

HyperMINE is a software tool which helps manipulate large amounts of satellite imagery data. It automatically reads and detects raw data downloaded from USGS (United States Geological Survey). It reads the metadata contained in the image files themselves or in the separate files that accompany, it orders and catalogues the information and presents the user with a structured form of that respective information.

The tool has modules for visualizing geo-referenced satellite imagery, database indexing and storing, complex database extraction, and image feature extraction and classification algorithms.

The system is developed in Java and has a multilayered and modular architecture, which eases future development. The main layers and modules are presented in Fig. 1.

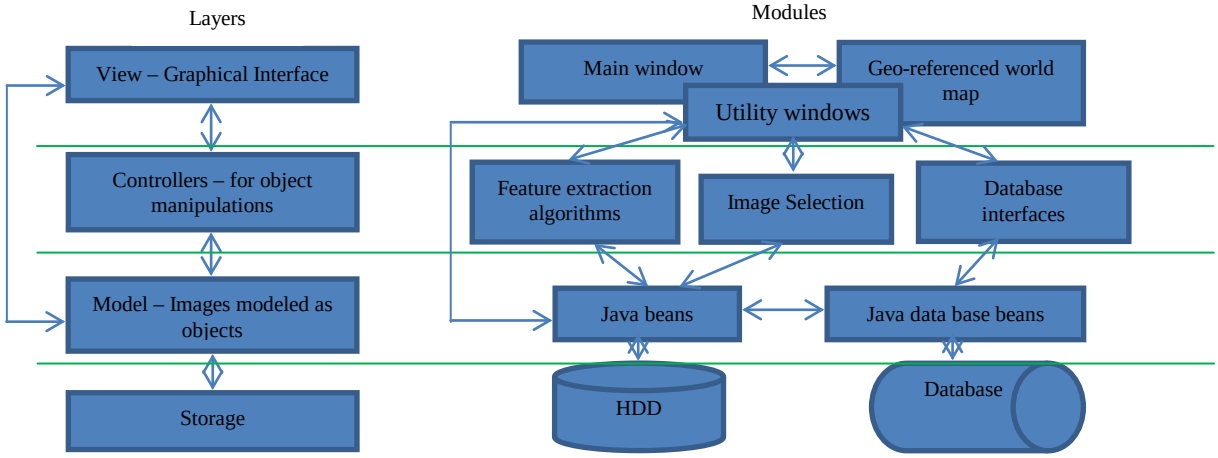

Fig. 1. GeoMIN system architecture by functionality layers and the modules associated to each layer

From a storage point of view, the application is capable of reading/writing information both from/on a mass storage device and from/in a database.

The database module for metadata information is developed in MySQL with B-Tree (Binary-tree) indexing and has the following tables, presented in Fig 2. The data hypercube is well represented as all the features (space, time, satellite as well as other characteristics) are present in the table. Indexing is made on the satellite, sensor, band fields in the *images* table and two association tables (*sensors_for_satellite* and *bands_for_sensor*) are used to rapidly populate and adapt the drop boxes used in the query window. This helps to make pre-selections before defining the final set of parameters for the final query, reducing the required query response time for an otherwise complex and time consuming query.

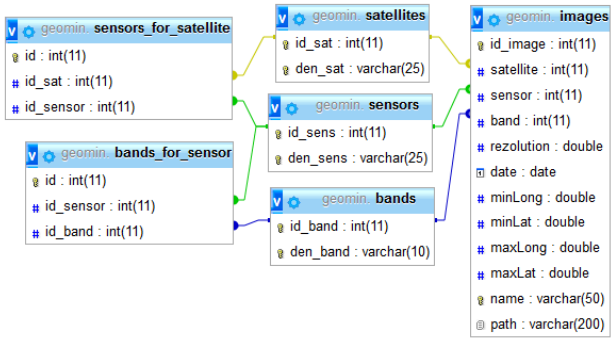

Fig. 2. MySQL database tables and relations used for storing metadata information

In order to facilitate the operations with the next two upper layers, java beans and java database beans (using the JPA technology) have been written. They represent a mean by which the application can map multiple sources of data into a common form making it easier to use different algorithms in the application.

The next layer of controllers contains the modules used to manipulate common data form produced by the model layer. Here we can find a module, image selection, used for different operations, such as sorting data and fast extraction from the collection of data objects loaded in RAM. In addition there is a module, database interfaces, used to ease the work with the database for insertion and extraction of information.

One of the more complex and which will know the greatest future development is the feature extraction algorithms module, which contains methods for extracting image features, data classification, and other image manipulations. The last two modules can also be seen as menu items in the main window presented in Fig. 3.

The top layer, the view layer which represents the interface with the user, is composed of one main window, a world view window and several other utility sub windows. The main window is where all the data can be viewed and ordered in different ways, and from where most of the systems functionalities can be accessed. The world view is a graphical interface from where geo-referenced images can be viewed, complex queries to the database can be made by selecting an interested geographical area or point, and where external data sources, such as other geo-referenced images and shape files can be added.

All these modules interconnect in order to aid the end user in the data mining process and in the extraction of the information hidden in the huge amounts of data.

IV. HyperMINE capabilities and applications

The main source for the data currently used in HyperMINE is the USGS database from where sets of images acquired during the Landsat 4, 5, 7 and 8 satellite missions have been downloaded. They cover the south part of Romania, with the area of interest being Bucharest. At the moment of writing the article the HyperMINE database contained the number of image acquisitions and corresponding characteristics presented in Table 1.

TABLE 1. SATELLITE DATA SOURCES AND CHARACTERISTICS

Source	Number of acquisitions	Number of bands	Resolution (m)	Period
Landsat 4	3	4	60	1989,1992
Landsat 5	58	7	30	1986-2011
Landsat 7	35	9	30	1999-2003
Landsat 8	24	12	30	2013-2015

As each band in the Landsat missions is stored as a separate image, the resulting total number of raw images is 1021 and requires almost 80GB of disk storage. In addition, each image was divided into 200x200 pixel patches, resulting in a total number of approximate 830.000 images. Each one has the metadata presented in Fig. 2 stored into the database. The resulting size of the database is of around 400 MB which needs to be loaded into RAM on the MySQL server, but together with the indexing made on most of the fields except image name and path assures fast responses to queries, as none of the complex queries goes over the 1s limit.

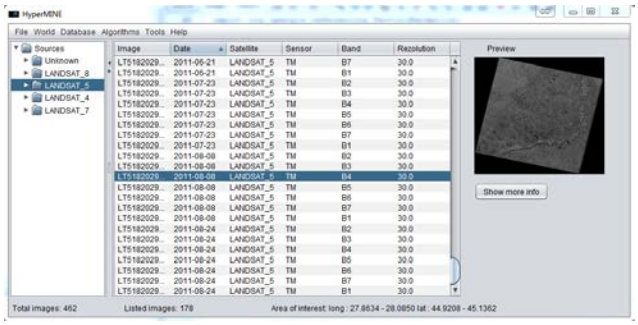

Fig. 3. Main window of HyperMINE

From the main window shown in fig. 3, the following actions can be made:

- Raw data (images) selection;
- Access the World view window;
- Work with the database;
- Apply different feature extraction and classification algorithms on the selected images;
- Use different tools that work with shape files;
- View different other information such as a quicklook 200x200 pixel image of the current selected image, total number of images found in the last query, the number of images that exist in the selected branch, and the current area of interest;

After adding raw data or after making a query to the database, the application auto populates the tree on the left of the main window with information in the following order: satellite, sensor (as branches) and bands (as leafs). Upon selecting a branch or a leaf the table in the center is populated automatically and instantly.

From the table in the center the information can be sorted and moved around to better suit the needs of the user. This is done by clicking on the table headers and by clicking and dragging around the headers.

HyperMINE uses libraries such as GeoTools which aids in the manipulation of geo-referenced images and data. The World view window is developed using this library. It mainly contains a geo-referenced map of the world having a resolution of 21600x10800 pixels and a shape file laid on top of the map image to give the borders for all the countries. From this widow, you can add/remove other geo-referenced sources, using the layers menu. From the Images menu queries can be made directly from the map by specifying a point or by drawing an area of interest. Upon doing this an utility window will pop up where more information can be added in order to fill the interested data characteristics such as the satellite source, the sensor, band and period of acquisition. After receiving the results for the query another fine selection can be made in the main window as presented in Fig. 3 to further inspect the images.

In Fig. 4 two possible visualization methods integrated in HyperMINE are presented for Landsat 5 images. The area is restrained to the geographical area previously defined on the World map. The four acquisitions are taken in 21-06-2011, 23-07-2011, 08-08-2011 and 24-08-2011. They are visualized in RGB format on left and NDVI (Normalized Difference Vegetation Index) values on the right. This is done by using bands 3, 2 and 1 (red, green and blue) to generate the RGB images and bands 4 (near infrared) and 3 for the calculation of NDVI.

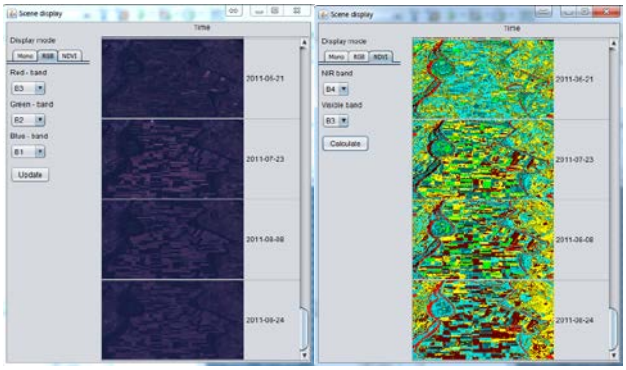

Fig. 4. RGB (left) and NDVI (right) representation of 4 Landsat 5 images from a preselected area in the Great Isle of Braila, Romania during June, July and August 2011

It can be easily observed from the SITS presented in Fig. 4 that vegetation (represented by turquoise and light green) changes throughout the summer season, as different crops such as wheat are harvested revealing the ground (represented by dark red).

The application momentarily uses just metadata in the database indexing process and supports only fast semantic queries, but further developments will be made in order to support CBIR (Content Based Information Retrieval) and widen the application area for HyperMINE.

V. Conclusions

HyperMINE presents itself as a fast spatio-temporal data mining system that aids in the manipulation of the hypercube of data resulted in the forming of SITS needed for different applications such as change pattern detection and classification.

The modular multilayer architecture permits it to rapidly gain new features and still maintain flexibility for future development.

What sets HyperMINE apart from the architectures presented in Chapter II is the combination of the architectures proposed in [1], [2], and [3]. Also the ability of spatially selecting the interested area for information extraction, the manipulation and fine selection of the data inserted in the algorithm processing is a big plus, reducing the workload of the user.

The application lacks CBIR methods presented in [4], [5] or [6], but this module will be present in a future development alongside with other classification and feature extraction algorithms.

Acknowledgment

The work has been funded by the Sectoral Operational Programme Human Resources Development 2007-2013 of the Ministry of European Funds through the Financial Agreement POSDRU/187/1.5/S/155536.

References

[1] Daniela Espinoza-Molina, Mihai Datcu, “Architecture Concept for Earth Observation Data Mining System”, IGARSS 2013, pp. 1729-1732.

[2] Muhammad Mazhar Ullah Rathore, Anand Paul, Awais Ahmad, Bo-Wei Chen, Bormin Huang, and Wen Ji, “Real-Time Big Data Analytical Architecture for Remote Sensing Application”, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.

[3] Paolo Mazzetti and Stefano Nativi, “Multidisciplinary Interoperability for Earth Observation: Some Architectural Issues”, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 5, no. 3, june 2012, pp. 1054-159.

[4] Gholamhosein Sheikholeslami, Wendy Chang, Aidong Zhan, “SemQuery: Semantic Clustering and Querying on Heterogeneous Features for Visual Data”, IEEE Transactions on Knowledge and Data Engineering, vol. 14, no. 5, september/october 2002, pp. 988-1002.

[5] Chi-Ren Shyu, Matt Klaric, Grant J. Scott, Adrian S. Barb, Curt H. Davis, Kannappan Palaniappan, “GeoIRIS: Geospatial Information Retrieval and Indexing System—Content Mining, Semantics Modeling, and Complex Queries”, IEEE Transactions on Geoscience and Remote Sensing, vol. 45, no. 4, april 2007, pp. 839- 852.S. Chen, B. Mulgrew, and P. M. Grant, “A clustering technique,” *IEEE Trans. Neural Networks*, vol. 4, pp. 570–578, July 1993.

[6] Daniela Espinoza-Molina, “Earth-Observation Image Retrieval Based on Content, Semantics, and Metadata”, IEEE Transactions on Geoscience and Remote Sensing, vol. 51, no. 11, November 2013, pp. 5145-5159.

[7] Lionel Gueguen, Mihai Datcu, “A Similarity Metric for Retrieval of Compressed Objects: Application for Mining Satellite Image Time Series”, IEEE Transactions on Knowledge and Data Engineering, vol. 20, no. 4, april 2008, pp. 562-575.

[8] Jean-Francois Im, Felix Giguere Villegas and Michael J. McGuffin, “VisReduce: Fast and Responsive Incremental Information Vizualization of Large Datasets”, IEEE International Conference on Big Data, 2013, pp. 25-32.