EARTH-OBSERVATION DATA ACCESS: A KNOWLEDGE DISCOVERY CONCEPT FOR PAYLOAD GROUND SEGMENTS

Daniela Espinoza-Molina and Mihai Datcu

German Aerospace Center, D-82234 Wessling - Germany

ABSTRACT

This paper proposes an alternative solution for enhancing the access to the data content. Our solution presents a knowledge discovery architecture concept, whose intention is to implement a communication channel between the Payload Ground Segments and the end-user who receives the content of the data sources coded in an understandable format associated with semantics and ready for the exploitation. The architecture concept is defined as a modular system composed of five modules allowing functions such as primitive feature extraction, tiling of the image content, metadata extraction, finding scenes of interest by content, enriched metadata, and semantics, the interpretation and understanding of the image content, semantic definition of the image content, and visualization of the image database via human machine interfaces. All these functionalities are integrated and supported by a relational database management system.

Key words: Earth-Observation images, data mining, knowledge discovery, visual data mining .

1. INTRODUCTION

In recent years the ability to store large quantities of Earth Observation (EO) satellite images has greatly surpassed the ability to access and meaningfully extract information from it. The state-of-the-art of operational systems for Remote Sensing data access (in particular for images) allows queries by geographical location, time of acquisition or type of sensor. Nevertheless, this information is often less relevant than the content of the scene (e.g. specific scattering properties, structures, objects, etc.). Moreover, the continuous increase in the size of the archives and in the variety and complexity of EO sensors require new methodologies and tools - based on a shared knowledge - for information mining and management, in support of emerging applications (e.g.: change detection, global monitoring, disaster and risk management, image time series, etc.).

Along the years, several solutions were presented for accessing the Earth-Observation archives as for example queries of the image archive using a small number of parameter like: geographical coordinates, acquisition times, etc. [WDS+09]. Later, the concept of query by example allowed to find and retrieve relevant images taking into account only the image content, provided in the form of primitive features, several systems following this principle appeared for instance [ACS99], [MD10], [DDP+03]. However, later the problems of matching the image content (expressed as primitive features) with semantic definitions adopted by human were evident; causing the so-called semantic gap [SWS+00]. With the semantic gap, the necessity of semantic definition was clearly demonstrated.

In this article we propose an alternative solution for enhancing the access to the data content. Our solution presents a knowledge discovery concept, whose intention is to implement a communication channel between the Payload Ground Segments (EO data sources) and the end-user who receives the content of the data sources coded in an understandable format associated with semantics and ready for the exploitation. The first implemented concepts were presented in Knowledge driven content based Image Information Mining (KIM) [DDP+03] and Geospatial Information Retrieval and Indexing (GeoIRIS) system as examples of data mining systems [SKS+07]. Our new concept is developed in a modular system composed of the following components 1) the data model generation implementing methods for extracting relevant descriptors (low-level features) of the sources (EO images), analysing their metadata in order to complement the information, and combining with vector data sources coming from Geographical Information Systems. 2) A database management system, where the database structure supports the knowledge management, feature computation, and visualization tools because of the modules for analysis, indexing, training and retrieval are resolved into the database. 3) Data mining and knowledge discovery tools allowing the end-user to perform advanced queries and to assign semantic annotations to the image content. The low-level features are complemented with semantic annotations giving meaning to the image information. The semantic description is based on semi-supervised learning methods for spatio-temporal and contextual pattern discovery. 4) Scene understanding counting on annotation tools for helping the user to create scenarios using EO images as for example change

Proc. 'ESA Living Planet Symposium 2013', Edinburgh, UK
9–13 September 2013 (ESA SP-722, December 2013)

detection analysis, etc. 5) Visual data mining providing Human-Machine Interfaces for navigating and browsing the archive using 2D or 3D representation. The visualization techniques perform an interactive loop in order to optimize the visual interaction with huge volumes of data of heterogeneous nature and the end-user.

2. KNOWLEDGE DISCOVERY ARCHITECTURE CONCEPT

The Earth-Observation data mining and knowledge discovery system intends to implement a communication channel between the EO data sources and the end-user who receives the content of the data coded in an understandable format associated with semantics.

The architecture concept and its components are described in Fig. 1. Here, it can be seen that the component are: 1) Data Model Generation (DMG), 2) Database Management System (DBMS), 3) Query, Data mining and Knowledge Discovery (KDD), 4) Interpretation and Image Understanding, and 5) Visual Data Mining (VDM). They are described in the next sub-sections.

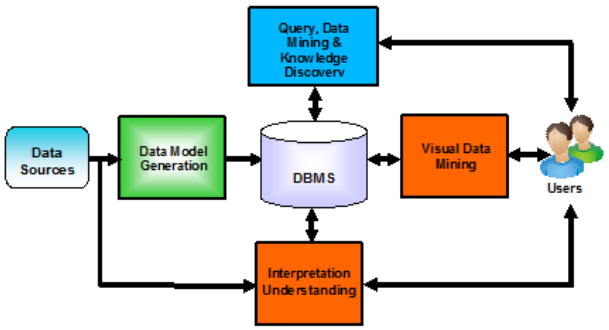

Figure 1. Architecture components of Earth Observation data mining and knowledge discovery system. All components interact with the database for creating a communication channel between the data sources and the end-user.

The starting point of the system operation is when a new image acquisition is done. This starts the Data Model Generation (DMG) module transforming the EO data from an initial form of full EO products to much smaller value added products, which includes image descriptors (primitive features), metadata, image patches, etc. All this information is stored into the database enabling the other system functionalities as for example queries, data mining, visualization, etc.

2.1. Data Sources

The inputs of the system are the different data sources, which are the EO images (Optical and Synthetic Aperture Radar (SAR)) with their associated metadata (i.e. acquisition time, incidence angles, geographical coordinates, etc).

2.2. Data Model Generation

The DMG focuses on the content and context analyses of the different EO data sources.

The Content Analysis of the DMG takes as input the Earth Observation products: the Earth Observation (EO) images, which are either optical or TerraSAR-X [DLR07], and their metadata (the xml files). The first input source is the Earth Observation image (Optical and TerraSAR-X scenes) and their processing parameters. The output is the image content descriptors in the form of vectors (e.g., feature vectors), which can later be used either for classification purposes or in the knowledge discovery framework. The second input is the metadata in form of annotations included in xml files or information in the header of geotiff files. The output is a set of descriptors. Finally, the further steps the EO image (raster data) will be completed with vector data in the form of lines, points or polygons (GIS information depending on it is available).

The context analysis is done by applying patch-based feature extraction method, thus the methods consider the whole patch as one entity, using the relations between all the pixels within the patch and not only one pixel information [AGD12]. It is important to note that the efficiency of the query, data mining and knowledge discovery depends on the robustness and accuracy of the image descriptors.

2.3. Database Management System

A relation Database Management System (DBMS) acts as the core of the system, interacting with all the components and supporting their functionalities. The use of a DBMS provides some advantages such as the natural integration of the different kinds of information, the ensuring of the relation integrity, the speed of the operations, etc.. The information generated by the DMG is mapped into a relational database. Thus, all the information about the EO image composed of tiles, coordinates, metadata, primitive features, quick-looks etc. are mapped into a table based scheme, which implements proper relations between the tables and indexes for optimization processes.

2.4. Query, Data Mining and Knowledge Discovery

This component is composed of two modules 1) query and 2) data mining and knowledge discovery. They have different functionality but they work together.

The *query module* allows the end-user finding desired scenes and retrieving images containing the required information. The queries are the starting point for the *data mining and knowledge discovery module (KDD)* module. The KDD module is finding hidden pattern in the image database and retrieving the relevant scenes according to query parameters. It requires running data mining methods in order to search into the entire image database and pick out the relevant images. Data mining might include methods for clustering and classification, similarity metrics, retrieving and ranking, etc. The image archive can be queried in the following ways:

1. Query based on metadata: Metadata entries such as geographical coordinates, acquisition angles and time, type of product, etc. are used as query parameters. The use of metadata can be used to create a complex scenarios as for example image time series by querying the available images of a zone within a time range, study of the incidence angle for determining available structures.

2. Query based on spatial content: The queries can use the implicit spatial information of the images given as geographical location (latitude and longitude coordinates). The queries based on location answer questions as for example the proximity between two objects (distances), the objects within a given boundary (containment), or objects in a given direction [EMD13].

3. Query based on semantics: In the queries based on standard metadata, the user is limited to the metadata entries for querying rich image archives. More advanced queries can be performed by including semantics to the image content. The use of semantics will help the user to better understand the image content. Using semantics, the end-user can enter a simple label i.e forest, urban area, etc. in the form a text or select from the available labels in the semantic catalogue to perform the query. Fig. 2 shows an example of query using semantic labels. In this case, all the patches annotated as *parking lots* were retrieved.

4. Query based on image content: The query by example or CBIR relies on *similarity metrics* computed between the query image passed as parameter and the images stored into the database. Later, the images are ranked according to these metrics. The image retrieval is based on the top ranked scenes. Fig. 3 shows an example of implementation [EMQD12]. This tool is implemented using compression based techniques in order to describe the image content and the Fast Compression Distance [CD12] as similarity metric. Fig.3 shows a patch containing *chaparral* is passed as query parameter and the system retrieves the 56 top ranked patches with chaparral.

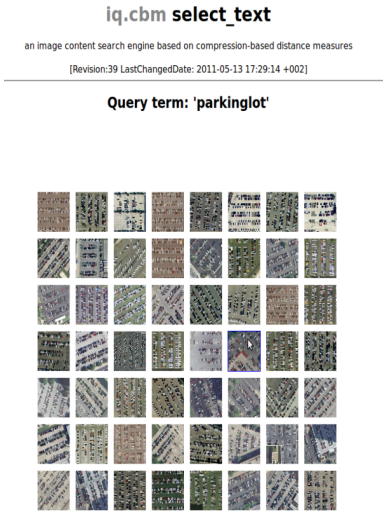

Figure 2. Example of query using semantics. The system retrieves all the patches annotated as parking lots.

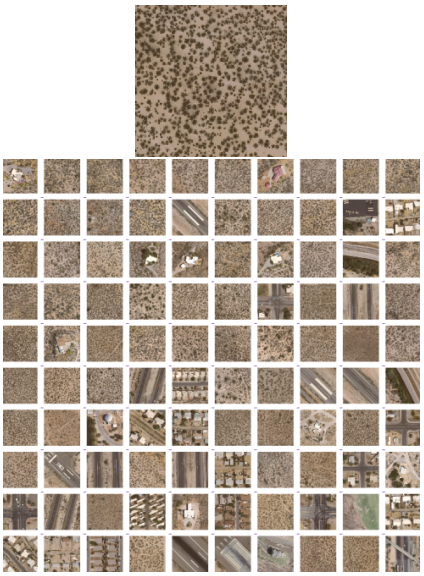

Figure 3. Example of based on image content. The end-user selects chaparral and the system passes it as query parameter, computes the similarity metrics, ranks and retrieves the results.

The previous queries used in combination are a powerful tool for exploring spatial patterns in the EO image database.

2.5. Interpretation and Image Understanding

While the data mining tools allow the user finding and retrieving a desired image from huge archives, this tool

allows the user finding objects of interest using one or few scenes. The image interpretation and understanding tool is based on semi-supervised learning methods complemented with relevance feedback tools. This tool is also used for labelling the image content by defining *semantics*.

Semantic definition is an arduous and expensive task that can be optimized using supervised or un-supervised machine-learning methods. Active learning algorithms are iterative sampling schemes where a classifier is adapted regularly by feeding it with new labelled samples. The key idea behind active learning is that a machine learning algorithm can achieve greater accuracy with fewer training labels if it is allowed to choose the data from which it learns [Set09].

These concepts can be implemented for labelling the image content or semantic definition as follows:

1. The training data set is obtained interactively from the Graphical User Interface (GUI). The training data set refers to a list of image patches marked as positive or negative examples by the end-user.
2. The training data set is passed to the classifier, which makes the prediction and retrieves the results.
3. The end-user judges the results and refines the training data. These steps can be supported by Relevance Feedback. The Relevance Feedback has a GUI allowing automatically ranking the suggested images, which are expected to be grouped in the class of relevance.
4. When the user is satisfied with the results, he stops the active learning loop and the new label is written into the DBMS catalogue.

During the active learning two goals are tried to achieve: 1) learn the targeted image category as accurately and as exhaustively as possible and 2) minimize the number of iterations in the relevance feedback loop.

2.6. Visual Data Mining

The Visual Data Mining component allows interactive exploration and analysis of very large, high complexity, and non-visual data sets stored into the database. It provides to the end-user an intuitive tool for data mining by presenting a graphical interface, where the selection of different images and/or image content in 2-D or 3D space is achieved through visualization techniques, data reduction methods, and similarity metrics to group the images. Actually, Visual Data Mining relies on powerful GUIs with functionalities such as browsing, querying, zooming, etc. enabling to navigating into the EO database. An example of Visual Data Mining displaying a set of hundred thousand TerraSAR-X image patches (size 160x160 pixels) is presented in Fig. 4.

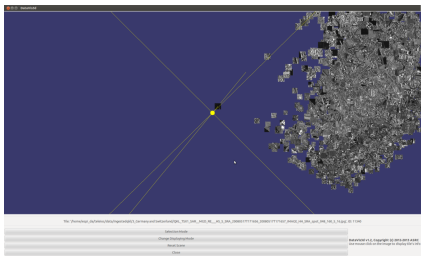

Figure 4. The Visual Data Mining tool displays the image database in a 3D space in order to help the en-user navigating and browsing the content.

Example of zooming-in function of Visual Data Mining using a collection of optical data is displayed in Fig. 5

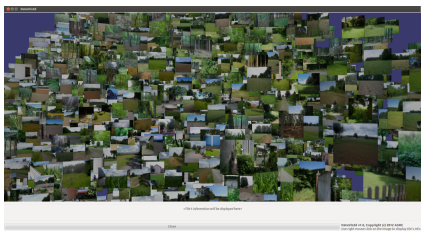

Figure 5. Zooming-in into a optical image collection using Visual Data Mining tool.

3. CONCLUSIONS

In this paper, we presented an architecture concept for knowledge discovery and data mining systems using Earth-Observation images. The architecture is presented as a modular system integrating several components with well-defined functionality. The main operation of the system starts with the ingestion of different EO data sources (i.e. TerraSAR-X, optical images) during the data model generation. The data model generation performs the tiling of the image content, the feature extraction based on tiles, the quick-looks generation, and the metatada extraction giving as result a complete model of the EO data, which later will be enhanced by adding semantic labels. Functions like data mining and knowledge discovery help the end-user in defining semantic of the image content and in finding hidden patterns in the image archive. In addition, visual data mining functions allow the end-user the exploration and exploration of huge image archives since it is based on advanced visualization techniques.

REFERENCES

[ACS99] Peggy Agouris, James Carswell, and Anthony Stefanidis. An environment for content-based image retrieval from large spatial databases. *ISPRS Journal*

of Photogrammetry and Remote Sensing, 54(4):263 – 272, 1999.

[AGD12] Popescu Anca, I. Gavat, and Mihai Datcu. Contextual descriptors for scene classes in very high resolution sar images. *IEEE letter on Geoscience and Remote Sensing*, 9(1):80 –84, jan. 2012.

[CD12] Daniele Cerra and Mihai Datcu. A fast compression-based similarity measure with applications to content-based image retrieval. *Journal of Visual Communication and Image Representation*, 23(2):293 – 302, 2012.

[DDP+03] M. Datcu, H. Daschiel, A. Pelizzari, M. Quartulli, A. Galoppo, A. Colapicchioni, M. Pastori, K. Seidel, P.G. Marchetti, and S. D'Elia. Information mining in remote sensing image archives: system concepts. *IEEE Transactions on Geoscience and Remote Sensing*, 41(12):2923 – 2936, dec. 2003.

[DLR07] DLR. *TerraSAR-X, Ground Segment, Level 1b Product Data Specification, TX-GS-DD-3307*. Dec 2007. http://sss.terrasar-x.dlr.de/pdfs/TX-GS-DD-3307.pdf.

[EMD13] Daniela Espinoza-Molina and Mihai Datcu. Earth-Observation Image Retrieval based on content, semantics and metadata. *IEEE Transactions on Geoscience and Remote Sensing*, agu 2013.

[EMQD12] Daniela Espinoza-Molina, M. Quartulli, and M. Datcu. Query by example in earth-observation image archive using data compression-based approach. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS) 2012*, pages 6035–6038, 2012.

[MD10] Inés María Gómez Muñoz and M. Datcu. System design considerations for image information mining in large archives. *Geoscience and Remote Sensing Letters, IEEE*, 7(1):13 –17, jan. 2010.

[Set09] Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[SKS+07] Chi-Ren Shyu, Matt Klaric, Grant J. Scott, Adrian S. Barb, Curt H. Davis, and Kannappan Palaniappan. GeoIRIS: Geospatial Information Retrieval and Indexing System-Content Mining, Semantics Modeling, and Complex Queries. *IEEE Transactions on Geoscience and Remote Sensing*, 45(4):839–852, 2007.

[SWS+00] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349 –1380, dec 2000.

[WDS+09] M. Wolfmuller, D. Dietrich, E. Sireteanu, S. Kiemle, E. Mikusch, and M. Bottcher. Data Flow and Workflow Organization- The Data Management for the TerraSAR-X Payload Ground Segment. *IEEE Transactions on Geoscience and Remote Sensing*, 47(1):44 –50, jan. 2009.