ANALYZING AND RETRIEVING REMOTE SENSING IMAGES FROM LARGE DATA ARCHIVES

Lorenzo Bruzzone[1], Begüm Demir[1], Francesca Bovolo[1], Carsten Brockmann[2], Norman Fomferra[2], Michele Iapaolo[3], Rajesh Jha[4], Jun Lu[4], Ralf Quast[2], Kerstin Stelzer[2], Luis Veci[4]

[1] Dept. of Information Engineering and Computer Science, University of Trento, Trento, Italy
[2] Brockmann Consult GmbH, Germany
[3] European Space Agency, ESRIN, Italy
[4] Array Systems Computing Inc., Canada

ABSTRACT

During the last decade, several optical and SAR sensors operate on board of satellites showing a variety of properties in terms of spatial, spectral, and temporal resolutions. Thus, millions of Earth Observation (EO) scenes are collected in very large EO data archives. Analyzing, mining and retrieving useful information from them is a big challenge. EO data archives grow rapidly motivating the need of efficient and effective processing and analysis tools. This will be evermore true when the new Sentinel missions will be launched and operated by ESA. In this context, this paper presents the activities developed in the framework of the ESA Long Term Data Preservation (LTDP)- Product Feature Extraction and Analysis project. This project aims to efficiently exploit the above-mentioned huge amount of data by defining: i) feature extraction methods for populating an EO data base with a set of effective features computed on different kinds of remote sensing data (i.e., SAR, optical and also time series of them), and ii) data analysis methods for extracting the semantic from the features in the context of different scenarios and applications.

Index Terms— information mining, remote sensing image retrieval, large remote sensing data archives.

1. INTRODUCTION

Nowadays, more and more satellites with optical and SAR sensors on-board have been launched due increased user demand, and the developments of satellite technology has increased the variety, amount, and resolution (spatial, spectral, and temporal) of Earth Observation (EO) data. Accordingly, millions of single-date as well as time-series of EO scenes have been acquired, resulting in very large EO data archives from which analyzing, mining and retrieving useful information are challenging [1]. EO data collections growing at a rapid rate motivates the need for efficient of effective tools to process and analyze the data. This situation will be further enriched when the new Sentinel missions will be launched and operated by ESA.

In order to efficiently exploit the already available huge amount of EO data and those that will be available with the Sentinel satellites, the Product Feature Extraction and Analysis (PFA) project aims at implementing two main parts: i) a feature extraction part that aims to effectively derive sets of features from different kinds of EO data, i.e., SAR, optical and also time series of SAR and optical images, and ii) a scenarios part that aims to analyze the features obtained in the first part in the framework of specific information extraction scenarios. Then, usefulness of scenarios will be demonstrated through "real-life" applications.

The high spatial and temporal resolution of images acquired by the new generation of satellite sensors (e.g., future Sentinel 1 and 2 missions for high spatial resolution SAR and optical images, respectively; and future Sentinel 3 mission for high temporal resolution images) require robust feature extractors that can emphasize the high information content of images. In the project we focus the attention on the implementation of feature extractors that can be effective on a large amount of EO data and on several kinds of EO data. These features extractors include: i) Features capable to effectively model the spatial/geometrical information in a large variety of EO image data (e.g. SAR and optical high resolution images) such as attribute filters (which contain as special case morphological filters) [2], attribute morphological profiles, etc. ii) Features that capture the multitemporal nature of satellite data such as the backscattering temporal variability and long-term coherence for SAR time series, the Fourier descriptors for optical and SAR time-series [3], etc. iii) Features that model the multiscale nature of spatial information in EO data such as

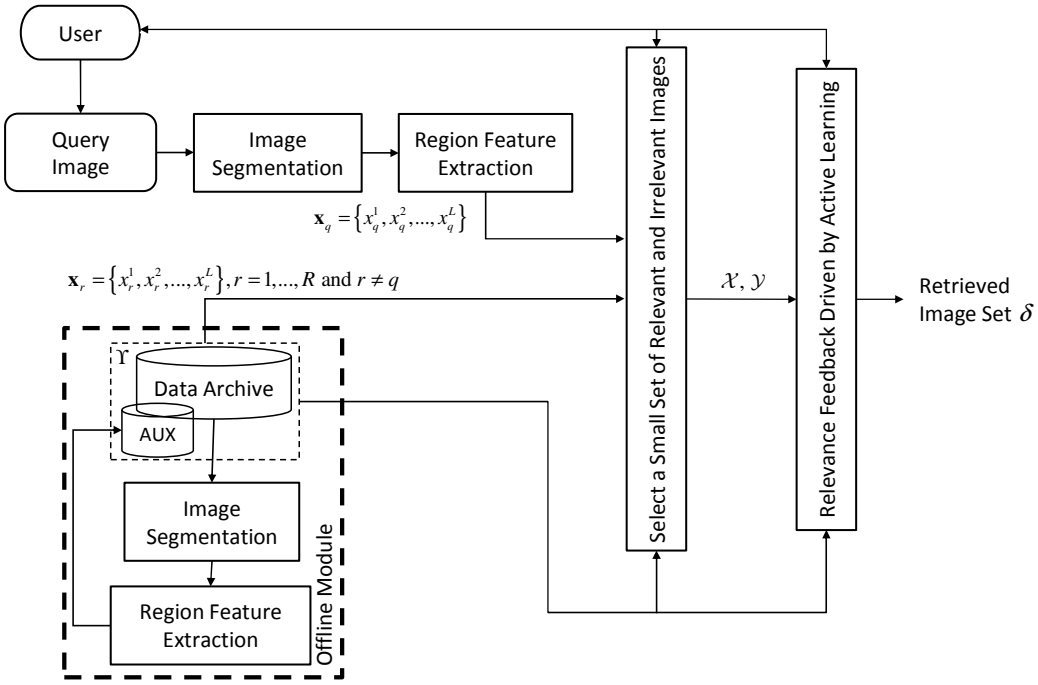

Figure 1: Flowchart of the first scenario

stationary and non-stationary discrete Wavelet transform and Gabor filters. These methods can be applied to both SAR and optical images, as well as to model the multiscale time variability of temporal signatures in time-series.

Extracted features will be employed in three main scenarios: i) content based image retrieval; ii) content based time-series retrieval; and iii) unsupervised classification with kernel methods. The first and second scenarios are devoted to fast and effective content based image retrieval on single images and image time-series, respectively. To this end a query data that can be an EO image, parameter values (such as a threshold values), temporal-trend of a time series or a step-change in bitemporal images should be defined. Once the query is fixed, efficient approaches are required to retrieve from large EO data archives images (or time series) that match the query. Here active-learning-based methods [4] are considered in the context of relevance feedback [4, 5] for both scenarios in order to efficiently exploit interaction with the user. The classification stage of the content based retrieval will be based on machine learning classifiers, such as Support Vector Machine (SVM) classifier, which are non-parametric (and thus suitable for any kind of data) and widely recognized as effective [6]. As in the analysis of EO archives it is not realistic to have ground truth data, the third scenario aims to search for the similar images to the query image without using any prior information on the archive. Here, the most promising unsupervised classification techniques, i.e., kernel based methods as kernel k-means [7], will be considered. It is due to the fact that i) it provides much more accurate classification results compared to standard techniques (i.e., k-means); it is distribution free (and thus can be applied to any kind of feature and thus of EO data, i.e., optical, SAR, and time series of SAR and optical data); and it shows superior performance when data classes have a non-linearly separable structure (as for many remote sensing images).

In the project special emphasis is devoted to: 1) give priority (when possible) to the development of data independent methodologies within the above-mentioned scenarios (i.e., methodologies that can be suitable for SAR and optical images, and time series of SAR and optical images); and 2) develop a coherent data processing framework where the methodologies being implemented for the individual scenarios can be exploited for the other considered scenarios.

The effectiveness of scenarios will be demonstrated through three applications: 1) retrieval of images with algal bloom; 2) retrieval of images urban area; and 3) retrieval of pairs of images in the archive that show the same kind of changes associated with burned areas in the forest.

2. SCENARIOS CONSIDERED IN THE PROJECT

This section is devoted to introduce the three scenarios to be implemented in this project. Let us consider an archive Υ made up of remote sensing images $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_R\}$ acquired from different remote sensing missions (i.e. both optical and SAR images belong to the archive) where R the number of images in the archive. $\mathbf{x}_i$ is the i-th image defined as

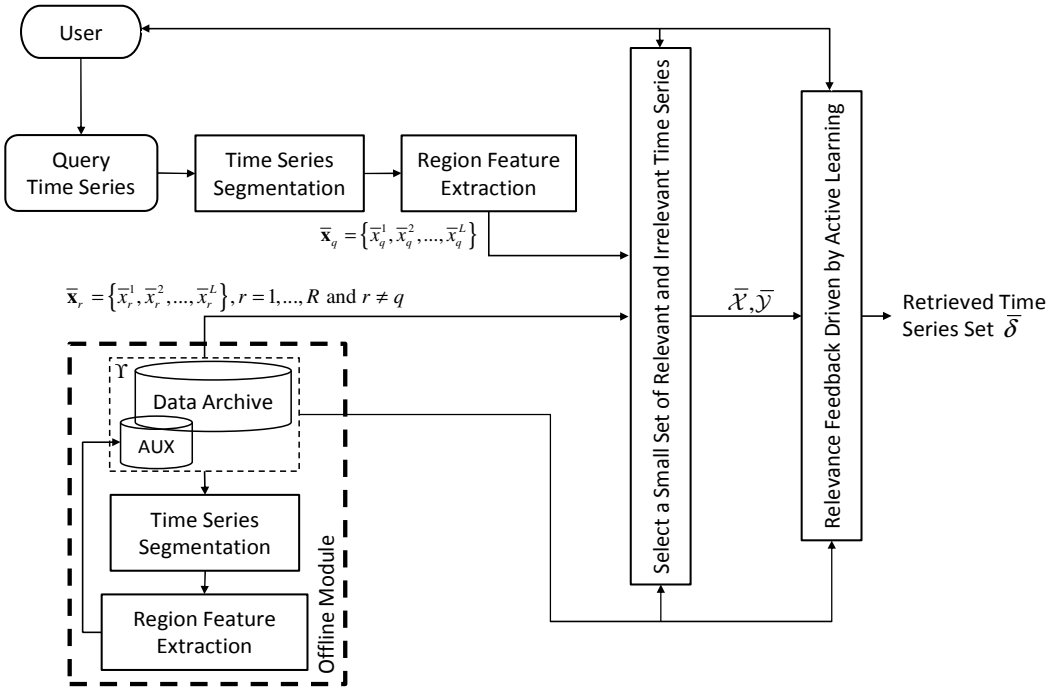

Figure 2: Flowchart of the second scenario

$\{x_i^1, x_i^2, ..., x_i^L\}$, $r = 1,..., R$, where x_i^l, $l = 1,..., L$ is the l-th feature computed at the feature extraction step. The primitive features can be computed on a pixel basis or based on segments. Usefulness of segmentation depends on the considered application. The archive Υ can be considered as a set of single images. Within the archive Υ image time series $\bar{\mathbf{x}}_i$ can be built by including images acquired on the same geographical area at different times. A time series $\bar{\mathbf{x}}_i$ is defined as $\{\bar{x}_i^1, \bar{x}_i^2, ..., \bar{x}_i^l\}$ where $\bar{x}_i^l$ can be one among the x_i^l single date features or a specific time series feature. Within such an archive content based retrieval can be performed both on images as well as time series. In the following descriptions of the steps to attain these goals are given.

A. *Content Based Image Retrieval*

The first scenario is devoted to search and retrieve images from the archive Υ, which are similar to the query image. Figure 1 shows a flowchart of the first scenario. A query $\mathbf{x}_q = \{x_q^1, x_q^2, ..., x_q^L\}$ is selected from the archive Υ having according to the end-user needs. In order to initialize the retrieval process an initial set $\mathcal{X}$ of relevant and irrelevant images (with respect to the query one) should be constructed. The user can decide to select these images either randomly or using a similarity measure. In the latter case, the similarity between the query image and images $\mathbf{x}_r = \{x_r^1, x_r^2, ..., x_r^L\}, r = 1,..., R,\ r \neq q$ in the archive is estimated by a certain metric. Then, a small number of images from the archive that have the highest similarity and highest dissimilarity to the query image is selected as an initial set $\mathcal{X}$ of relevant and irrelevant images, respectively. Accordingly, image retrieval is modeled as a two-class problem: one class includes relevant images, and the other one consists of the irrelevant ones. The set of labels relevant and irrelevant is indicated with $\mathcal{Y}$. After obtaining the initial set $\mathcal{X}$, relevance feedback (RF) procedure driven by Active Learning (AL) is performed. AL iteratively expands the number of annotated images by selecting the most informative images from the data archive for annotation [4]. The main goal of AL is to identify the image in the archive that, when annotated in an interactive way, one can optimize the search in the archive. At the convergence of RF, a set $\delta = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_U\}, U << R, \delta \subset \Upsilon$ of the most relevant images from the archive Υ is selected [4].

B. *Content Based Time Series Retrieval*

The second scenario is devoted to search and retrieve time series from the archive Υ, which are similar to the query time series. A query $\bar{\mathbf{x}}_q = \{\bar{x}_q^1, \bar{x}_q^2, ..., \bar{x}_q^L\}$ is selected from the archive Υ having according to the end-user needs. Figure 2 shows a block scheme of the second scenario. In order to initialize the retrieval process an initial set $\bar{\mathcal{X}}$ of relevant and irrelevant time series (with respect to the query one) should be constructed. The user can decide to select these

images either randomly or using a similarity measure. In the latter case, similarity between the query time series and time series $\bar{\mathbf{x}}_r = \{\bar{x}_r^1, \bar{x}_r^2, ..., \bar{x}_r^L\}, r = 1, ..., R, \ r \neq q$ in the archive is estimated by a certain metric. Then, a small number of time series from the archive that have the highest similarity and highest dissimilarity to the query time series is selected as an initial set $\bar{\mathcal{X}}$ of relevant and irrelevant time series, respectively. Accordingly, retrieving time-series is modeled as a two-class problem: one class includes relevant time-series, and the other one consists of the irrelevant time-series. The set of labels relevant and irrelevant is indicated with $\bar{\mathcal{Y}}$. After obtaining the initial set $\bar{\mathcal{X}}$, RF procedure driven by AL is performed. At the convergence of RF, a set of $\bar{\delta} = \{\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, ..., \bar{\mathbf{x}}_u\}, u << G, \bar{\delta} \subset \Upsilon$ the most relevant time-series from the archive is selected.

C. *Unsupervised Classification with Kernel Methods*

The third scenario is devoted to retrieve the images from the archive Υ that are similar to the query image $\mathbf{x}_q$ using an unsupervised classification method. This scenario aims to search for the similar images to the query image without using any prior information on the archive (i.e., unsupervised). Figure 3 shows a flowchart of the third scenario. Let $\mathbf{x}_q = \{x_q^1, x_q^2, ..., x_q^L\}$ be a query image selected by the user. Initially an unsupervised classification method is applied to all the images in the archive Υ. Then, the closest cluster to the query image is found. The closest cluster can be found by estimating the distances between each cluster center and the query image $\mathbf{x}_q$ in the feature space. Then, the set of images located in the closest cluster to the query image $\mathbf{x}_q$ is selected as relevant images; whereas those belong to the other classes are chosen as irrelevant images. It is worth noting that if the query image $\mathbf{x}_q$ is already included in the archive, due to applying clustering to all the images in the achieve, the set of images assigned to the same cluster with the query image $\mathbf{x}_q$ is directly selected as relevant images; whereas those belong to the other classes are chosen as irrelevant images. In other words, there is no need to estimate the closest cluster.

3. CONCLUSION

This project addresses emerging methods and tools for data product features and information extraction, in view of possible implementations for enriching data description and easing the use of archived data. We are elaborating and implementing a number of EO data exploitation scenarios whose usefulness will be demonstrated by means of "real-life" applications: 1) retrieval of images with algal bloom; 2) retrieval of images urban area; and 3) retrieval of pairs of images in the archive that show the same kind of changes associated with burned areas in the forest.

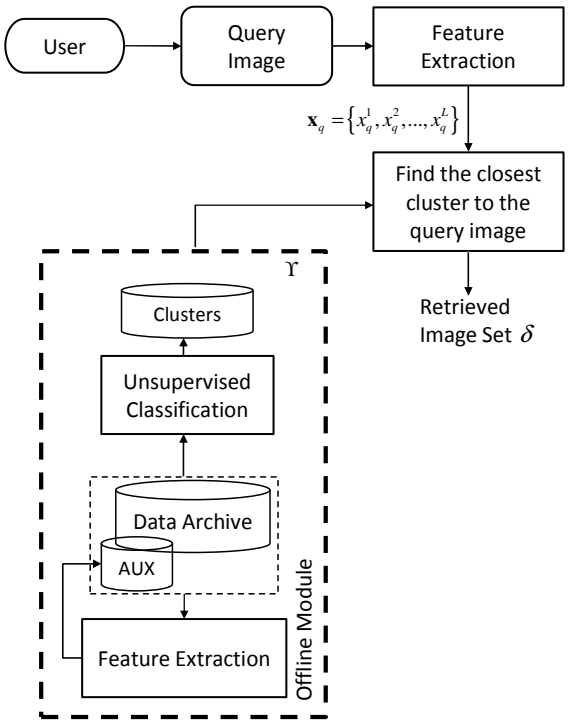

Figure 3: Flowchart of the third scenario

REFERENCES

[1] M. Datcu, S. D'Elia, R. L. King, and L. Bruzzone, "Introduction to the special section on image information mining for earth observation data," *IEEE Trans. Geosci. Rem. Sen.*, vol. 45, no. 4, pp. 795–798, 2007.

[2] M. Dalla Mura, J.A. Benediktsson, B. Waske, L. Bruzzone, "Morphological attribute profiles for the analysis of very high resolution images," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 10, pp. 3747-3762, 2010.

[3] L. Bruzzone, M. Marconcini, U. Wegmuller, A. Wiesmann, "An advanced system for the automatic classification of multitemporal SAR images*," IEEE Trans. Geosci. Rem. Sens.*, vol. 25, no. 13, 2004, pp. 1491-1500.

[4] B. Demir, L. Bruzzone, "An Effective Active Learning Method for Interactive Content-Based Retrieval in Remote Sensing Images", *International Conference on Geoscience and Remote Sensing Symposium*, Melbourne, Australia, 2013.

[5] M. Ferecatu, N. Boujemaa, "Interactive Remote-Sensing Image Retrieval Using Active Relevance Feedback", *IEEE Trans. on Geosci. and Rem. Sensing*, vol. 45, no. 4, pp. 818-826, 2007.

[6] G. Camps-Valls and L. Bruzzone, "Kernel-based methods for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 6, pp. 1351–1362, 2005.

[7] I.R. Zhang, A.I. Rudnicky, "A Large scale clustering scheme for kernel k-means," *IEEE Int. Conf. on Pattern Recog.,* pp. 289-292, 2002.