Complexity Based Analysis of Earth Observation Images

Authors: Daniele Cerra(2), Alexandre Mallet(1), Lionel Gueguen(1), and Mihai Datcu(1,2)
(1)GET/Télécom Paris, 46 rue Barrault, 75013 Paris, France
(2)German Aerospace Center (DLR), Oberpfaffenhofen, 82234 Weßling, Germany

Frame
Complexity based methods have been recently applied to image analysis, reaching good results in several domains: their main advantage is the independency from statistical models used by classical methods, models which are difficult to establish on very large, diverse and irregular datasets as in the case of satellite imagery. In complexity based analysis these limitations are discarded by establishing typical dictionaries and similarity measures, directly learned from the data, which don’t require any a priori knowledge.

These methods result in several applications in the field of satellite data:
- Classification
- Segmentation
- Indexing
- Artifact Detection
- Data mining of spatial-temporal patterns in Image Time Series

Examples of these applications are reported below.

Complexity based analysis is relatively new and has to be still fully understood, and its model-independency and high reliability in capturing patterns in any kind of data suggest that algorithms based upon it will be able to shed a light on many cases of satellite imagery. In complexity based analysis these limitations are difficult to establish on very large, diverse and irregular datasets as in the case of satellite imagery. In complexity based analysis, these methods result in several applications in the field of satellite data:

- Classification
- Segmentation
- Indexing
- Artifact Detection
- Data mining of spatial-temporal patterns in Image Time Series

Examples of these applications are reported below.

Information Theoretical Frame

Shannon Entropy (1950)

\[ H(X) = - \sum p(x) \log p(x) \]

Probabilistic Complexity of a random source \( X \)

\[ I(X, Y) = H(X) | Y - H(Y) \]

Mutual Information: Classic definition for two random sources \( X \) and \( Y \).

Link Shannon-Kolmogorov

\[ H(X) = \sum p(x) K(x | p) + O(1) \]

Kolmogorov Complexity (1956)

\[ K(x) = \min_{p(x)} |p(x)| \]

Algorithmic Complexity of a dataset \( x \), regarded as the shortest program \( p \) that outputs \( x \) and halts on an universal Turing machine

\[ I(x : y) = K(x) + K(y) - K(x, y) \]

Algorithmic Mutual Information between two datasets \( x \) and \( y \).

Complexity Based Similarity Measures

From the definition of algorithmic mutual information it is possible to derive a similarity metric, the Normalized Compression Distance (NID), between two datasets \( x \) and \( y \):

\[ NID(x, y) = \frac{K(x, y) - \min\{K(x), K(y)\}}{\max\{K(x), K(y)\}} \]

The NID is approximated with the similarity metric NCD (Normalized Compression Distance) in order to be computable and is used in all the applications described below.

Complexity Based Similarity Measures

From the definition of algorithmic mutual information it is possible to derive a similarity metric, the Normalized Compression Distance (NID), between two datasets \( x \) and \( y \):

\[ NID(x, y) = \frac{K(x, y) - \min\{K(x), K(y)\}}{\max\{K(x), K(y)\}} \]

The NID is approximated with the similarity metric NCD (Normalized Compression Distance) in order to be computable and is used in all the applications described below.

Artifact Detection with Compression Based Similarity Metrics

Images may contain blemishes or artificial structures resulting from their processing or directly from the sensors. These artifacts decrease the data quality and can lead to analysis and interpretation problems. Compression based similarity metrics are able to detect these defects automatically.

Workflow

- Scheme for artifact detection: artifact free elements are 32 × 32 windows manually selected. A complexity comparison is applied with compression based similarity measures in order to build a significant feature space. A decision is then taken over the feature space which provides the detection map, through a classification or a simple thresholding.

Joint database encoding and index creation. First, each object is coded independently with an optimal two part coder. Then, a distance matrix based on NCD (Normalized Compression Distance) is computed between each pair of objects. Using an hierarchical agglomerative clustering on this distance leads to several data volume partitions. Finally, based on the Minimum Description Length criterion, we select the partition which compresses the most the whole data volume. The resulting index is included in the code.

Image Presenting Artifacts

Images containing artifacts

Artificial free image elements

Complexity comparison

Detection

Classification

Workflow

A two part coder is used to compress spatio-temporal events contained in Satellite Image Time Series (SITS). This allows to derive an optimal index of SITS events, enabling queries based on information content.

Mineral of Satellite Image Time Series (SITS)

A two part coder is used to compress spatio-temporal events contained in Satellite Image Time Series (SITS). This allows to derive an optimal index of SITS events, enabling queries based on information content.

Image Self-Indexing and Classification

Workflow. After image segmentation, each region of the image is matched with representative datasets for each class using the similarity metric NCD. Each region is considered to belong to the class which minimizes the NCD.

Classification. In the example, a mosaic of SPOT subsets. From upper-left corner in clockwise order clouds, forest, city, sea. The urban area, which is not homogeneous, is correctly assigned by the algorithm to the same class.

Segmentation. The image is coded as an undirected graph to improve segmentation performance.

Indexing of TerraSAR-X Imagery by Hierarchical Clustering. Compression techniques can also be applied to SAR images. To obtain the results above, a NCD distance matrix has been computed between each pair of images (belonging to a set of 10x10 subsets). Afterwards, a binary tree is fit to the data in order to cluster subsets with similar characteristics.

Results

This sequence of size 600x600x9 is a subseries of the encoded series of size 2000x1500x35. The whole series contains about 25000 spatio-temporal patterns. The magenta rectangle represents the query, and the orange squares represent the query answer. In the series, we detect the roads which are covered by snow.