# Complexity Based Analysis of Earth Observation Imagery: an Assessment

Daniele Cerra, Alexandre Mallet, Lionel Gueguen, and Mihai Datcu

***Abstract*—this paper proposes complexity based analysis as a valid alternative to classic image analysis methodologies for Earth Observation imagery, which are heavily dependant on the assumed data models. We will show the power of this totally model-free, data-driven methods by presenting three very different applications relying on complexity based analysis: image classification, artifact detection, and database compression which enables queries directly on the compressed content.**

***Index Terms*— Data Compression, Data Mining, Kolmogorov Complexity, Image classification.**

## I. Introduction

COMPLEXITY based image analysis offer a valid alternative to classic Bayesian or maximum-likelihood methods, which are widely used in several domains like segmentation, classification or inference clustering. A major drawback of Bayesian methods is that they require strong a priori knowledge of the data, which is possible on well described kinds of objects, but may restrict the efficiency of these methodologies on Earth Observation images databases: in fact, the large and steadily growing volumes of data provided by satellites, together with the large variety, diversity and irregularity of the observed scenes, make hard to establish enough general statistical description models for the data.

These limitations are overcome in the analysis methods described in this paper, which aim at extracting information and computing similarities on the sole basis of the data complexity, estimated with solutions based on classical information theory, or by means of typical dictionaries directly learned from the data. The main advantage of this approach is that it is totally data-driven, being independent of any statistical model. Thus, classical concepts of information theory may gain new meaning in various fields of Earth Observation imagery analysis, thanks to these recent concepts.

The paper is structured as follows. Section II presents a reminder on classical and algorithmic information theory concepts, presented in parallel, while following sections present practical applications of these concepts within the field of image analysis. Section III will be related to classification of both optical and SAR Earth Observation data, performed by a hierarchical clustering obtained on the basis of a complexity based similarity measure. Section IV will present artifact detection in optical remotely sensed images: it is shown how the concepts of complexity enable one to highlight blemishes introduced during the image formation process. Section V presents how to build an index that can be queried related to a compressed Image Time Series database. Finally, section VI reports some conclusions and future perspectives.

D. Cerra and M. Datcu are with the German Aerospace Center (DLR), 82234, Wessling, Germany (email: {daniele.cerra, mihai.datcu}@dlr.de).

M. Datcu, A. Mallet and L. Gueguen are with the Paris Institute of Technology (ENST), 46 rue Barrault, 75013 Paris, France

## II. Information Theoretical Frame

Complexity issues can be addressed from a probabilistic point of view (Shannon) [1] or from an algorithmic point of view (Kolmogorov) [2]. Their basic ideas are different, but Shannon's classical approach can be linked to Kolmogorov's concept of computational complexity [3]: this section will introduce this relation, presenting a summary of the main recent works on these topics.

In all the following definitions and theorems, we will assume $X$ to be a random variable with a set of possible outcomes $x$ and an associated probability distribution $P(X = x) = f(x)$.

### A. Shannon entropy and Kolmogorov complexity

Both concepts of Shannon entropy and Kolmogorov complexity aim at measuring the quantity of information contained in a binary string.

*Definition 1*: The entropy $H(X)$ of the random variable $X$ is given by:

$$H(X) = -\sum_{x} f(x) \log f(x) \tag{1}$$

This definition can be interpreted as the length in bits needed to encode the outcomes of $X$, which can be obtained, for example, through the Shannon-Fano code; nevertheless, such approach related to probabilistic assumptions does not provide the informational content of individual object and their possible regularity: that lacuna is filled by the Kolmogorov complexity that evaluates an intrinsic complexity for any isolated object, independently of any description formalism.

*Definition 2*: The Kolmogorov complexity $K(x)$ of a string $x$ is the length of the shortest program $q$ that outputs $x$ and halts on an appropriate universal machine, such as an universal Turing Machine, being defined as:

$$K(x) = \min_{q \in Qx} |q| \ , \qquad (2)$$

with $Qx$ being the set of programs generating $x$ and $|q|$ the length of the program $q$. So, strings presenting recurring patterns have low complexity, while the complexity of random strings is high and almost equals their own length. Such coding interpretation of the Kolmogorov complexity is confirmed by the fact that $m(x) = 2^{-K(x)}$ defines a universal distribution for a priori model of $x$ [4]. It is important to remark that $K(x)$ is not a computable function of $x$.

### B. Mutual information in Shannon and Kolmogorov

An important issue of the informational content analysis is to be able to estimate how much information an object contains about another one. From Shannon's probabilistic point of view, the solution is brought through the mutual information.

*Definition 3*: The mutual information $I(X,Y)$ is defined in classical Shannon information theory as:

$$I(X,Y) = H(X) - H(X \mid Y) = H(Y) + H(X) - H(X,Y), \quad (3)$$

where $X$ and $Y$ are two random variables, $H(A \mid B)$ is the *conditional entropy* of $A$ given $B$ and $H(A,B)$ is the *joint entropy* of $A$ and $B$.

*Definition 4*: In the Kolmogorov complexity frame, the algorithmic mutual information between $x$ and $y$ is:

$$\mathrm{I}_W(x:y) = K(x) - K(x \mid y) = K(x) + K(y) - K(x,y), \quad (4)$$

defined up to an additive constant, where the conditional Kolmogorov complexity $K(x \mid y)$ of $x$ related to $y$ is the length of the shortest program to compute $x$ if the string $y$ is given as an auxiliary input to the computation, while the joint complexity $K(x,y)$ is the length of the shortest program which outputs $x$ followed by $y$.

### C. Link between the two theories

In spite of the fundamental differences between the two concepts of Shannon entropy and Kolmogorov complexity, a link between them has been established, stated in the following theorem [3].

*Theorem 1*: If $f$ is a probability mass function associated to a random source $X$, then:

$$0 \le \left( \sum_x f(x) K(x) - H(X) \right) \le K(f) + O(1). \quad (5)$$

This means that the sum of the expected Kolmogorov complexities of all the outcomes $x$ of a random variable $X$ equals the Shannon entropy of $X$, up to an additive constant. Likewise, it is established that the expected algorithmic mutual information equals the probabilistic mutual information up to an additive constant. The results presented above result in the following equality:

$$H(X) = \sum_x f(x) K(x \mid f) + O(1), \qquad (6)$$

which enables one to assimilate the conditional complexity $K(x \mid f)$ to the uncertainty of the probabilistic distribution $-\log f(x)$. The importance of this result lies in the fact that it constitutes a simple estimation way of the conditional complexity under the a priori knowledge of a probabilistic distribution.

### D. Compression-based similarity measures

The Normalized Information Distance (NID) between two objects $x$ and $y$ is the length of the shortest program that computes $x$ knowing $y$, as well as computing $y$ knowing $x$, i.e. it is proportional to the quantity $\max\{K(x \mid y), K(y \mid x)\}$ [5]; the distance becomes, after normalization:

$$NID(x,y) = \frac{K(x,y) - \min\{K(y), K(x)\}}{\max\{K(x), K(y)\}} \ . \qquad (7)$$

$NID(x,y)$ is a metric, with $NID(x,y) = 0$ *iff* $x = y$ and $NID(x,y)$ = 1 meaning maximum distance between $x$ and $y$.

Since Kolmogorov complexity is non-computable, so it is the $NID$. To find a suitable approximation of $K(x)$, it can be stated to represent the length of the shortest lossless compressed file $x^*$ obtained compressing $x$: $K(x)$ represents then a lower bound for what a real compressor can achieve.

This allows approximating $K(x)$ with $C(x)$, i.e. the length of the compressed version of $x$ obtained with a standard lossless compressor $C$ such as Gzip. The equation (7) becomes then the Normalized Compression Distance $NCD(x,y)$, and can be explicitly computed as

$$NCD(x,y) = \frac{C(x,y) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}} \ , \qquad (9)$$

with $C(x,y)$ representing the size of the file obtained by compressing the concatenation of $x$ and $y$. It is proved that the conditions for *NCD* (defined as an *admissible distance*) to be a metric hold under certain assumptions [6].

It has to be remarked that this metric is very general and experiments have been carried out to compute distances within any kinds of data: among those simple text files, music samples [6], sample dictionaries from different languages, DNA samples [7].

Recently [8], a direct link has been found between this metric and other distance measures and classification

techniques based on pattern matching and dictionaries extraction, such as Pattern Recognition based on Data Compression (PRDC) [9]: this suggests that many research perspectives are still open within this field.

## III. EO Images Hierarchical Clustering

The compression-based similarity measures introduced in the previous section are a powerful tool to discover similarities within satellite data with a total data-driven, model-free approach; furthermore, NCD has been recently shown to be noise-resistant [10].

To test how well can NCD perform in recognizing similar objects, we have tested its power on two totally different satellite imagery datasets: the first contains 60 SPOT 5 image subsets, single band and equally divided in 6 classes, and another containing 44 Synthetic Aperture Radar (SAR) TerraSAR-X subsets taken over Egypt, with Equivalent Number of Looks (ENL) equal to 4, and divided in 4 classes.

For each database, the NCD has been computed between each pair of objects in order to generate a distance matrix. To compare the overall distances the tool *maketree*, which is part of the tools provided by the open-source utilities suite Complearn [11], is then used to perform a hierarchical clustering, generating the best-fitting binary trees related to each distance matrix.

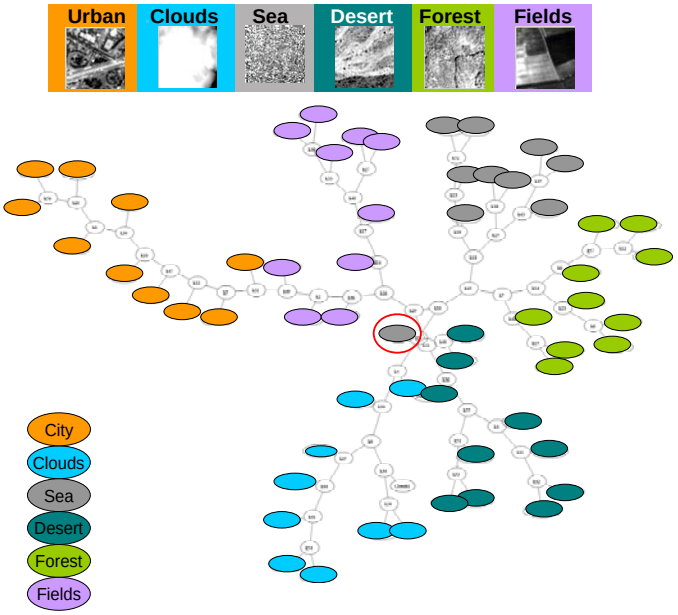


Fig.1. Visual description of the classes used (top) and hierarchical clustering of NCD values (bottom) applied to 60 SPOT 5 optical satellite images of size 64x64. The only false alarm is circled in red.

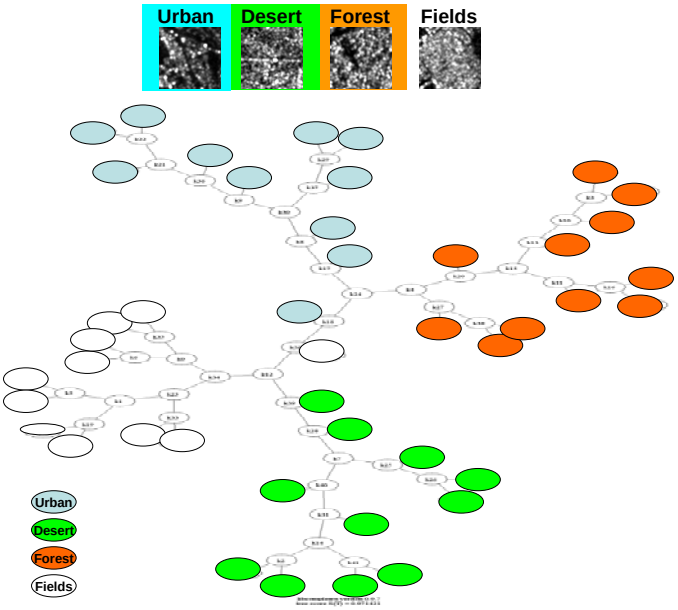


Fig.2. Visual description of the classes used (top) and hierarchical clustering of NCD values (bottom) applied to 44 TerraSAR-X images of size 64x64 with Equivalent Number of Looks (ENL) equal to 4.

Results in Figs. 1-2 show that all classes are well separated in both optical and SAR datasets, with only one “false alarm” in the first example. This confirms the discrimination power of NCD, which performs equally well on these two extremely different kinds of data: this is achieved without using any reference model or a priori knowledge of the data.

## IV. Image Artifacts Detection

EO images may contain blemishes or artificial structures introduced in the processing step or coming directly from the sensors (ref. top line in Fig.4): these artifacts decrease the quality of the images and can lead to analysis and interpretation problems; in [12] some steps are taken for their automatic detection. A visual analysis of such artifacts suggests that they alter the local complexity within the images, resulting in areas with complexity either too low or too high: therefore, a complexity comparison method may be able to detect these defects, under the assumption that it is possible to identify some artifact-free elements within each image.

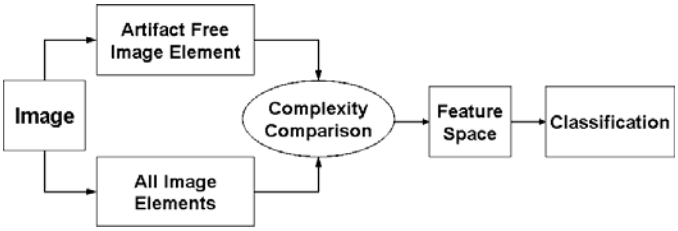


Fig.3. Workflow for artefact detection: all image elements (size 4 × 4) are compared with artefact free elements (size 32 × 32) manually selected. A complexity comparison is carried out by applying the NCD to build a significant feature space. A decision is then taken over that feature space which provides the detection map, through a classification or a simple thresholding.

The workflow is reported in Fig. X: the image elements are compared with NCD to the elements of the image without

artifacts, building a feature vector for each element. Classification can then be applied to the feature space to detect the artifacts.

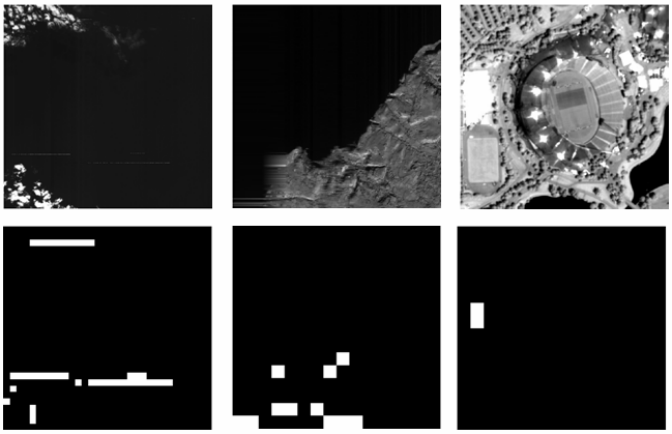

Fig.4. Results of artifact detection. On top, from left to right: SPOT image presenting the drop out artifact, infrared SPOT image presenting a degradation over the bottom left corner due to electronics failures in the sensor, and IKONOS saturated image. Bottom, from left to right: detections applied on images and manually thresholded. The inconsistent areas are highlighted. In the last case (bottom right) only the biggest saturated area is detected, because of the size of the windows employed in the algorithm.

In the examples presented in Fig.4, the output of the complexity comparisons is then clustered with the k-means algorithm and manually thresholded to output the artifact detection. Results are promising and show the adaptability of this methodology in detecting different kinds of artifact, confirming any model of the various artifacts to be unnecessary using this approach.

## V. Mining Satellite Image Time Series

In [13] a method is proposed to build an index of the content of a compressed Satellite Image Time Series (SITS) database. To achieve this, both optimal lossy and lossless source coding of the database are used; the general concept of indexing by content is not totally rejected, but it is adapted to perform a source coding of the database aimed at extracting objectively the information content: as clustering is used for lossless coding, it is possible to control the information loss with objective criteria. Therefore, the index is contained in the resulting code and it is equivalent to a dictionary.

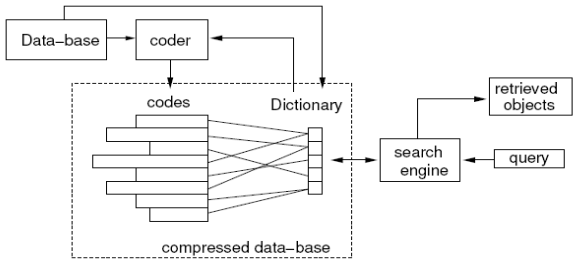


Fig.5. The figure presents the concept for building a compressed indexed data-base with additional compression.

The approach is presented in Fig. 5. First, a dictionary is computed from the database; then, a coder using this dictionary is able to code efficiently each object of the database, using the informational similarity measure and thus taking into account the inter-objects correlations. The resulting database representation is composed of a dictionary and the coded objects: this idea comes from the two-part representation enunciated by Rissanen [ref] and Kolmogorov [ref]. This coding scheme is lossless; however, the dictionary itself, which includes and information content index, is a lossy representation of the database: it contains the minimal sufficient information, according to Kolmogorov, to discriminate the data-volume objects. When the database is queried, only the information related to the dictionary is taken into account.

| Queries | total retrieved | false detections | missed detections | Precision | Recall |
|---|---|---|---|---|---|
| 1 | 38 | 23 | 6 | 0.39 | 0.71 |
| 2 | 48 | 3 | 29 | 0.93 | 0.60 |
| 3 | 72 | 19 | 16 | 0.73 | 0.76 |
| 4 | 94 | 18 | 39 | 0.80 | 0.66 |
| 5 | 64 | 15 | 20 | 0.76 | 0.71 |
| 6 | 352 | 167 | 50 | 0.52 | 0.78 |
| 7 | 114 | 61 | 45 | 0.46 | 0.54 |
| 8 | 79 | 14 | 25 | 0.82 | 0.72 |
| 9 | 79 | 40 | 35 | 0.49 | 0.52 |
| 10 | 137 | 20 | 40 | 0.85 | 0.74 |
| *average* | nd | nd | nd | 0.67 | 0.67 |

Table 1. Precision/Recall values for a set of 10 spatio-temporal patterns queries over a feature space of 4.500 clusters.

Some experiments have been carried out on the ADAM dataset [14], provided by the French Space Agency. The images, constituting the SITS, have been acquired by the satellites SPOT 1, 2 and 4, and have a resolution of 20m; the SITS comprises 38 images of size 3000×2000, and each image contains 3 spectral bands.

Table 1 shows Precision/Recall values [15] for the retrieval of a test set of 10 spatio-temporal patterns, queried on a database of about 25,000 spatio-temporal patterns extracted from an ITS, grouped in 4,500 clusters. As the data-base is not labeled, the false and missed detection are measured visually. Despite the high subjectivity of these results, the Recall and Precision averages of 0.67 are obtained.

The compression of the SITS database with the proposed method achieves then two goals: it compresses in a lossless way the images with a ratio of approx 1:3, and at the same time it enables query on the compressed database content with an acceptable Precision-Recall score.

## VI. Conclusions

This paper presented a recent approach to image analysis based on complexity estimations, and showed three different applications based on these concepts; complexity is estimated with dictionaries directly learned from the data and with data compression techniques and these methodologies rely heavily on classical and algorithmic information theory. In such approach there is no need of any a priori knowledge of the data, and satellite imagery varying greatly in content, resolution, and also sensor-type, may be analyzed with the same tools: this opens many doors into the field of image information mining.

Furthermore, in this field of research still many perspectives are open: the connections between algorithmic complexity, classical information theory and coding with dictionaries still are not completely clear and have to be precisely defined; this may lead to major changes in the standard analysis and processing chains of satellite imagery, since it could be possible to combine in a single step data compression and feature extraction.

## REFERENCES

[1] C. E. Shannon, *A Mathematical Theory of Communication*, Bell System Technical Journal, Vol. 27, pp. 379–423, 623–656, 1948.
[2] M. Li and P.M.B. Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications*, Springer, 1997
[3] P. Grünwald and P.M.B. Vitányi, *Shannon Information and Kolmogorov Complexity*, Amsterdam, the Netherlands, Sept. 2004.
[4] R. Solomonoff, *The Universal distribution and machine learning*, The Kolmogorov Lecture, Feb. 27, 2003, Royal Holloway, Univ. of London. The Computer Journal, Vol. 46, No. 6., 2003.
[5] M. Li, X. Chen, X. Li, B. Ma and P.M.B. Vitányi, *The Similarity Metric*, IEEE Trans Inf Theory, vol. 50, No. 12, pp. 3250-3264, Dec. 2004.
[6] R. Cilibrasi and P.M.B. Vitányi, *Clustering by Compression*, IEEE Trans Inf Theory, 51|4:1523-1545, 2005.
[7] M. Li et al., *An Information-based Sequence Distance and its Application to whole Mitochondrial Genome Phylogeny*, Bioinformatics 2001, 17(2):149-154.
[8] D. Cerra and M. Datcu, *A Model Conditioned Data Compression Based Similarity Measure*, in DCC, Snowbird, Utah, USA, Mar. 2008, p. 509.
[9] Watanabe T., Sugawara K., Sugihata H., *A New Pattern Representation Scheme Using Data Compression*, IEEE Trans Pattern Analysis Machine Intel, 24:579-590, 2002.
[10] M. Cebrian, M. Alfonseca, and A. Ortega, *The Normalized Compression Distance is Resistant to Noise*, IEEE Trans Inf Theory, 53|5:1895-1900, 2007.
[11] Complearn tool by R. Cilibrasi, A. Cruz, S. de Rooij, and M. Keijzer, based on the research of Cilibrasi, Vitányi, and Li. It is available at http://www.complearn.org/index.html.
[12] A. Mallet and M. Datcu, *Complexity Based Image Artefact Detection*, in DCC, Snowbird, Utah, USA, Mar. 2008, p. 534.
[13] L. Gueguen and M. Datcu, *A Similarity Metric for Retrieval of Compressed Objects: Application for Mining Satellite Image Time Series*, IEEE Trans Knowl Data Eng, vol. 20, No. 4, pp. 562-575, April 2008.
[14] ADAM data set: copyright CNES, 2000-2003. The set is composed of 57 SPOT images, freely accessible at http://medias.obs-mip.fr/adam .
[15] J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel, *Performance measures for information extraction*, Proc. DARPA Broadcast News Workshop, Herndon, VA, February 1999.