ABSTRACT

This paper presents a method which extracts information from Satellite Image Time Series which are new type of data set acquired with remote sensing technologies. The method is based on Multi-Information Bottleneck theory. The principle of inference for data clustering and clusters number selection is presented. Finally, the paper concludes showing examples presenting an information extraction from Satellite Image Time Series.

1. INTRODUCTION

Recently, the growing number of satellite image sensors has led to the acquisition of a huge amount of data. Processing techniques are needed in order to exploit this very informative material. Moreover, images of the same scene can be acquired several times a year because of the increasing number of satellites. Thus a new type of data sets can be created. In order to create Satellite Image Time-Series (SITS), a registration technique is used on several acquisitions of the same scene. The high spatial resolution of the sensors give access to detailed spatial structures, which are extended to spatio-temporal structures considering the time evolution of the scene. In order to exploit this huge amount of data, characterization of spatio-temporal patterns is essential. For example in a SITS, growth, maturation or harvest of cultures can be observed. State of the art tools for information extraction in SITS have been elaborated such as change detection, monitoring or validation of physical models. However, these technics are dedicated to specific applications [1]. Consequently in order to exploit the information contained in SITS, more general analyzing methods are required. Some methods for low resolution images and uniform sampling have been studied in [2]. For high resolution and non-uniform time-sampled SITS, new spatio-temporal analyzing tool is presented in [3, 4]. It is based on a Bayesian hierarchical model of information content. The concept was first introduced in [5, 6, 7] for information mining in remote sensing image archives. In a first stage, the extraction of information is data driven. Data is objectively represented. Usually, unsupervised methods are used to achieve this task. In a second stage, the extraction is user driven. Data is subjectively represented under the constraints provided by a user. In fact, the subjective representation is obtained from the objective representation by machine learning methods. The advantage of such a concept is that it is free of the application specificity and adapts to the user’s query. Due to the large amount of information contained in SITS, the quantity of information required to represent data is a crucial point. This paper addresses the problem of representing objectively and shortly the information contained in SITS by unsupervised clustering. From a compression point of view, clustering is equivalent to vector quantization. The method proposed produces a short length representation able to characterize spatio-temporal structures. In this paper, we propose a method based on Multivariate Information Bottleneck in order to estimate the optimal number of clusters and characterize spatio-temporal structures. In order to detect or recognize spatio-temporal patterns, it is essential to characterize information in a low-dimensional space. Features are extracted by fitting parametric models to data. This task can be viewed as a Bayesian hierarchical model in two stages. The first level of inference is the model fitting and the second level is the model selection. Then, an unsupervised clustering is processed on features space, reducing the complexity for fast retrieval of similar patterns. As clustering is equivalent to vector quantization, the problem can be viewed as a Rate-Distortion optimization. There is a trade-off between the amount of relevant information extracted (distortion defined with a divergence measure $d$) and the complexity of representation (rate expressed with the mutual information $I$). In order to determine this trade-off, we propose a criterion based on the Rate-Distortion curve. The paper is organized as follows. Section 2 introduces the Information and Multi-Information Bottleneck principles. Section 3 presents the relevant information that can be extracted from SITS. In Section 4 we give the Multi-Information Bottleneck principle for spatio-temporal structures characterization. Experiments and discussion are detailed in Section 5. Finally, Section 6 concludes the paper.

2. INFORMATION BOTTLENECK PRINCIPLE

The following sections present the theory of Information and Multi-Information Bottleneck. In the following sections, upper-case letters are used for random variables, while lower-case letters are used for realizations of random variables.

2.1 Information Bottleneck theory

Information Bottleneck emerged from Rate-Distortion theory. The problem is stated as follows: we would like a rel-
event quantizer $\tilde{X}$ to compress $X$ as much as possible under the constraint of a distortion measure between $X$ and $\tilde{X}$. In contrast, we also want to capture as much of information in $\tilde{X}$ as possible about a third variable $Y$. In fact, we pass the information that $X$ provides about $Y$ through a bottleneck formed by the compact summary formed by $\tilde{X}$. The problem is mathematically expressed as:

$$\min_{p(\tilde{x}|x)} I(\tilde{X},X) - \beta I(\tilde{X},Y)$$  \hspace{1cm} (1)

The algorithms for solving the problem are described in [8] and are mainly inspired from the Blahut-Arimoto algorithm [9]. They make the assumption of the following Markov chain: $Y \leftrightarrow X \leftrightarrow \tilde{X}$. However, Banerjee demonstrated in [10], that Information Bottleneck can be viewed as a Rate-Distortion problem based on the Bregman divergence. He considered $Z = p(Y \mid X)$ and $\tilde{Z} = p(Y \mid \tilde{X})$ as sufficient statistics for $X$ and $\tilde{X}$, respectively. $Z$ takes values over the set of conditional distributions $\{p(Y \mid x)\}$, and $\tilde{Z}$ takes values over the set of conditional distributions $\{p(Y \mid \tilde{x})\} = \tilde{Z}$. Therefore, the problem equivalent to Bottleneck Information is written as:

$$\min_{\tilde{Z}, p(\tilde{z}|z)} I(Z,\tilde{Z}) + \beta E_{Z,\tilde{Z}} [d(Z,\tilde{Z})]$$  \hspace{1cm} (2)

where $d$ is a Bregman Divergence, that corresponds here to the Kullback-Leibler divergence.

$$d(z,\tilde{z}) = \sum_y p(y \mid x) \log \frac{p(y \mid x)}{p(y \mid \tilde{x})}$$  \hspace{1cm} (3)

Cover and Thomas gave the solution to this problem for a fixed $\tilde{Z}^i$ [11].

$$p(\tilde{z} \mid z) = \frac{p(\tilde{z})}{N(z, \beta)} e^{-\beta d(z,\tilde{z})} \hspace{1cm} (4)$$

$$N(z, \beta) = \sum_{\tilde{z}} p(\tilde{z}) e^{-\beta d(z,\tilde{z})} \hspace{1cm} (5)$$

where $N(z, \beta)$ is the partition function. For fixed probabilistic assignments $p(\tilde{z} \mid z)$, the solution is given by:

$$\tilde{z} = E_{Z \mid z}[Z] \hspace{1cm} (6)$$

$$= \sum_{\tilde{z}} p(\tilde{z} \mid z)z \hspace{1cm} (7)$$

Using this two properties, Banerjee proposed in [10, 12] an iterative algorithm to compute $\tilde{Z}^i$ and $p(\tilde{z} \mid z)$. This algorithm is used to solve the problem, and to reach a local optimum of the functional. Finally, from this optimization the divergence $D_\beta$ and the rate $R_\beta$ can be computed with the following formulas:

$$D_\beta = \sum_{z,\tilde{z}} p(z)p(\tilde{z} \mid z)d(z,\tilde{z}) \hspace{1cm} (8)$$

$$R_\beta = \sum_{z,\tilde{z}} p(z)p(\tilde{z} \mid z) \log \frac{p(\tilde{z} \mid z)}{p(\tilde{z})} \hspace{1cm} (9)$$

![Figure 1: Markov graph representing the dependencies between variables. $Z_i$ are the conditional probabilities between the variables $X$ and $Z_i$.](image)

### 2.2 Multi-Information Bottleneck Theory

The Multi-Information Bottleneck principle generalizes the principle presented in the previous section and was first introduced in [13]. In this section we take a special case of this general principle by considering a set of independent variables $\{Y_i\}$ which contain the relevant information. As in the previous section, we define dependencies between variables with a Markov graph, shown in Figure 1. The bottleneck is represented in the Figure 2. Then, the Multi-Information Bottleneck problem is expressed as:

$$\min_{\tilde{Z}, p(\tilde{z}|z)} \sum_I I(Z_i,\tilde{Z}) + \beta E_{Z_i,\tilde{Z}} [d(Z_i,\tilde{Z}_i)]$$  \hspace{1cm} (10)

The set of Lagrangian parameters $\{\beta_i\}$ trades off between the compression and the relevant information extracted. In fact, the variables $\{Y_i\}$ enable to qualify the information contained in $X$ extracted from $X$ while the mutual information $I(X, X)$ quantifies this information. Therefore, the Lagrangian parameters control the qualification of the information. We consider the variables $Z_i = p(Y_i \mid X)$ and $\tilde{Z}_i = p(Y_i \mid \tilde{X})$. Each $Z_i$ takes values over the set of conditional distributions $\{p(Y_i \mid x)\} = \tilde{Z}_i$. Then, the problem can be expressed as:

$$\min_{\tilde{Z}_i, p(\tilde{z}_i|z)} \sum_{I} I(Z_i,\tilde{Z}_i) + \beta E_{Z_i,\tilde{Z}_i} [d(Z_i,\tilde{Z}_i)]$$  \hspace{1cm} (11)

Solutions to (10) can be explicitly calculated for two conditions, as in the equations (4) and (6). For fixed $\tilde{Z}_i$, solutions are:

$$p(\tilde{z} \mid x) = \frac{p(\tilde{z})}{N(x, \{\tilde{\beta}_i\})} e^{-\Sigma \beta(d(z,\tilde{z}_i))} \hspace{1cm} (12)$$

$$N(x, \{\tilde{\beta}_i\}) = \sum_{x} p(\tilde{x}) e^{-\Sigma \beta d(z,\tilde{z}_i)} \hspace{1cm} (13)$$

$$p(\tilde{z}) = \sum_{x} p(\tilde{z} \mid x)p(x) \hspace{1cm} (14)$$

For fixed probabilistic assignments $p(\tilde{z} \mid x)$, solutions are given by:

$$\tilde{z}_i = E_{X_i \mid Z_i}[Z_i] \hspace{1cm} (15)$$

$$= \sum_{x} p(x \mid \tilde{x})z_i \hspace{1cm} (16)$$

Finally, using these two properties we use an algorithm inspired from the one proposed in [10, 12]. These algorithms are similar to the Expectation-Maximization algorithm with the maximization step (12), (14) and the expectation step...
The Multi Rate-Distortion curve can be computed with the following formulas. We can define multi distortion depending on the variables \( Y_r \).

\[
D'_{\{\beta_j\}} = \sum_{x,z} p(x)p(\tilde{x} | x)d(z_i, \tilde{z})
\]

\[
R_{\{\beta_j\}} = \sum_{x,z} p(x)p(\tilde{x} | x)\log \frac{p(\tilde{x} | x)}{p(\tilde{x})}
\]

### 2.3 Optimal trade-off

In the algorithm of Information Bottleneck, the number \( k \) of \( \tilde{z} \) obtained after optimization is constrained by \( \beta \). Therefore, the initial number \( k \) is chosen to be equal to the number of realizations \( z \). Then, the optimal \( \beta \) corresponds to the maximum of the Rate-Distortion curve \( D(R) \) which is a parametric function of \( \beta \). The optimal \( \beta \) maximizes the second derivative of \( D(R) \) (19). This criterion has been experimentally studied in [14]. In fact we try to localize on a slow decreasing of distortion with the rate, which means that compression gains are not really noticeable.

We extend the principle to the multi Rate-Distortion curves by maximizing the Laplacian.

\[
\{ \hat{\beta}_j \} = \arg \sup_{\beta} \Delta D(R)
\]

\[
= \arg \sup_{\beta} \sum_i \frac{\partial^2 D'_{\beta_i}}{\partial R_{\beta}}
\]

In addition, local maxima are also points of interest. They determine the hierarchical structure of clusters. A natural cluster tree can be derived, by selecting clusterings obtained at each local maximum. A curve representing the Laplacian, where some local maxima exist, is drawn in Figure 5.

### 3. RELEVANT INFORMATION CONTAINED IN SITS

We want to characterize information contained in Satellite Image Time Series. Specialists qualify three types of information contained in satellite images: textural, geometrical and spectral information. These characterizations are considered to be independent. Therefore, by applying the Multi-Information Bottleneck principle to those information types, one characterizes the information contained in SITS. A problem is to find the variables that could contain relevant information. Consequently, we propose to characterize texture by Gaussian-Markov Random Field parameters and we characterize the spectral information by the spectral signature. For example, information is described with Gaussian Mixture in [15]. Geometrical information is not taken in consideration in this study.

#### 3.1 Gauss-Markov Random Field

Gauss-Markov Random Fields (GMRF) are parametric models which have presented interesting properties for characterizing textures in satellite images [16, 17]. We can extend the principle to a 3-dimensional signal. The field is defined on a rectangular grid. Let \( X_r \) be the signal, \( s \) belonging to a lattice and let \( N \) be the half of a symmetric 3-d neighborhood (Fig.3). GMRF are defined as follows:

\[
X_s = \sum_{r \in N} \theta_s(X_{s+r} + X_{s-r}) + e_s
\]

where \( e_s \) is a white Gaussian noise. Then parameters \( \Theta \) and the noise variance \( \sigma \) are estimated by Least Mean Squares, which corresponds to the Maximum Likelihood estimation considering a white Gaussian error. The equation (21) is expressed vectorially (22), by introducing a matrix \( G \) expressed with the values of the vector \( X \). Hence, the estimated parameters are expressed in the following equations.

\[
X = G\Theta + E
\]

\[
\Theta = \left(GG^T\right)^{-1}G^TX
\]

\[
\sigma^2 = X^TX - (G\Theta)^T(G\Theta)
\]

We denote the texture variable by \( T = (\Theta, \sigma) \). We calculate the estimate of parameters for each realization of the variable \( X \). Then, these estimates constitute a set of parameters \( \Omega_T \). The random variable \( T \) takes its value in the set \( \Omega_T \). Finally, the conditional probabilities \( p(T | X) \) are estimated...
with Bayes rules and the Gaussian distribution of the noise $E$ calculated with the equation (22). We assume that the parameters $\Theta, \sigma$ are equally distributed and $N$ is the length of $X$.

$$p(T \mid X) = \frac{p(X \mid T)}{\sum_{\Omega_T} p(X \mid T)}$$ (25)

$$p(X \mid T) = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\frac{1}{2} \frac{E^T E}{\sigma^2}}$$ (26)

### 3.2 Spectral information

Each image of the SITS is composed of three spectral bands. We model spectral information by a Gaussian model which is fully determined by its mean and variance. We denote by $S = (\mu_S, \sigma_S)$ these quantities which take values in the set $\Omega_S$. Similarly to texture parameters, spectral parameters are estimated for each realization of $X$. Finally, the conditional probabilities are computed using the following equations:

$$p(S \mid X) = \frac{p(X \mid S)}{\sum_{\Omega_S} p(X \mid S)}$$ (27)

$$p(X \mid S) = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\frac{1}{2} \frac{(X - \nu)^T (X - \nu)}{\sigma^2}}$$ (28)

### 4. MULTI-INFORMATION BOTTLENECK APPROACH FOR CLUSTERING

This section explains how to use previous results with the Multi-Information Bottleneck to calculate a soft clustering. We take into account two types of information, knowing that they are qualified to be textural and spectral. We estimate the parameters $T, S$ for each $X$ by maximum likelihood. Then, we define $\Omega_T, \Omega_S$ to be the sets that contain all the estimated parameters. In consequence, we can evaluate the conditional probabilities expressed by (25), (27) and denoted as follows:

$$z_1 = p(T \mid x)$$ (29)

$$z_2 = p(S \mid x)$$ (30)

Finally, using recursively the equations (12), (14) and (15), the algorithm described in [10] converges to a local minimum and gives the soft clustering $p(X \mid T)$. In order to find the optimal trade-off, we run the algorithm with varying parameters $\beta_1, \beta_2$. The methodology is represented in Figure 4.

Two channels of communication are clearly represented. It shows that information is data-driven extracted by two channels of communications before being fused in a single representation.

### 5. EXPERIMENTS AND DISCUSSION

For our experiments, we have worked on SITS provided by the CNES. Each image is composed of 3 spectral bands and has the size of $3000 \times 2000$ pixels. The series is composed of 38 images which represent the evolution of the countryside at the South East of Bucarest during one year. Moreover, the series is non uniformly sampled in time. We worked on a subseries of size $70 \times 70 \times 10$. A parallelepipedic partition of the data is done and we consider each parallelepiped as a realization of a random variable $X$. The partition is determined by the size (width $\times$ height $\times$ time) of the parallelepips which is also called the analyzing window size. We take a window of $10 \times 10 \times 5$. For the computation time problem, we have chosen to run the algorithm for several equal trade-off parameters ($\beta_1 = \beta_2$). Figure 5 shows the Laplacian of Rate-Distortion curves obtained with several trade-off parameters. There are several local maxima which indicate the existence of a hierarchical structure. Then, the number of clusters obtained at the optimal trade-off is 131. In addition, the Rate-Distortion curves give a way to quantify the extracted information. In our case, the information extracted can be encoded at 3.44 bits per symbol. Figure 6 represents one cluster in the data space. The cluster is representative of a spatio-temporal pattern given by an oblic line and a white part which disappear in time.

### 6. CONCLUSION

Nowadays only a few SITS exist, therefore the data type we considered is quite recent. However, with the increasing expansion of satellites, the number of SITS will grow. A new technique for information extraction has been presented in this paper. 3D texture models and spectral models have been extended for characterizing spatio-temporal structures. The method enables to find the number of classes contained in SITS by determining the critical number of clusters in the feature space. Finally, the method enables to quantify and qualify the extracted information.
REFERENCES


