BAYESIAN LEARNING ON GRAPHS FOR REASONING ON IMAGE TIME-SERIES

Patrick Héas* and Mihai Datcu†

*IRIT, Research Institute in Computer Science, Toulouse, France
Patrick.Heas@enseeiht.fr
†DLR, German Aerospace Center, Oberpfaffenhofen, Germany
Mihai.Datcu@dlr.de

Abstract. Satellite image time-series (SITS) are multidimensional signals of high complexity. Their main characteristics are spatio-temporal patterns which describes the scene dynamics. The information contained in SITS was coded using Bayesian methods, resulting in a graph representation [2].
This paper further presents a concept of interactive learning for semantic labeling of spatio-temporal patterns present in SITS. It enables the recognition and the probabilistic retrieval of similar events.
Graphs are attached to statistical models for spatio-temporal processes, which at their turn describe physical changes in the observed scene. Therefore, user-specific semantics attached to spatio-temporal events are modeled using combinations of parameters of a distance model between sub-graphs. Thus, the learning step is performed by the incremental definition of a spatio-temporal event type via user-provided positive and negative sub-graph examples. From these examples we infer probabilities of the Bayesian network, based on a Dirichlet model, that links user interest to a specific similarity measurement. According to the current state of learning, sub-graph posterior probabilities are estimated. Experiments, performed on a multitemporal SPOT image time-series, demonstrate the presented reasoning concept.

INTRODUCTION

During the last decades, the imaging satellite sensors have acquired huge quantities of data enabling the elaboration of SITS. However, our capability to store large volume of data has highly exceeded our capability to extract and interpret the relevant information. Therefore, SITS information mining systems are needed to bridge the semantic gap between information extracted from temporal and pictural multidimensional data, and user-specific meanings of available information in the same data. To cover such a broad problem, we establish an information flow between SITS content and user interest by modeling hierarchically the information content in SITS. On the first levels of the hierarchical modeling, strong families of models are applied to extract information using inference based on Bayesian and entropic methods. This unsupervised modeling results in a graph representation coding the information content of SITS [2]. The inferred graph $\mathscr{G}$ characterizes cluster trajectories in the dynamic feature space related to multitemporal (MT) objects.

In this paper we focus on a very important step in providing content-based query techniques: the interaction with the user and the flexible incorporation of user-specific

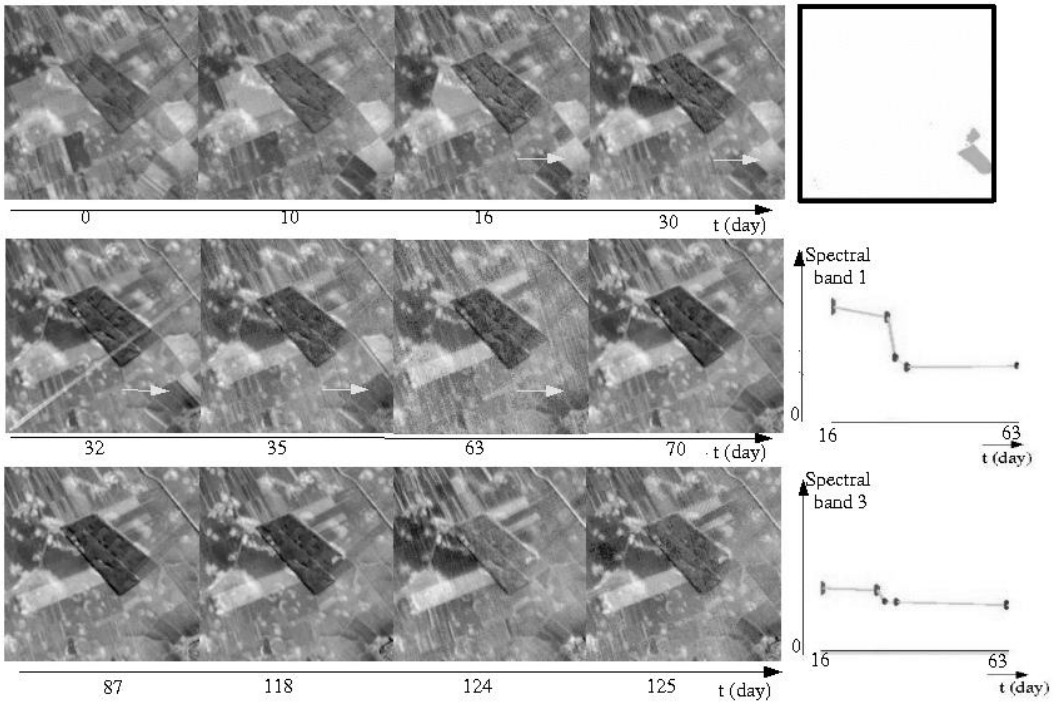

FIGURE 1. (Left) Example of a spatio-temporal pattern occurring in SITS which is associated with an harvest semantic $\mathscr{A}_v$. (Right) Corresponding MT class and cluster trajectory $\mathscr{G}_k$ characterizing the phenomenon.

interests. It constitutes the last level of the hierarchical information modeling.
Spatio-temporal processes present in a given time and spatial window of the STIS can possess subjective user-specific semantics denoted by $\mathscr{A}_v$ (e.g. harvests, cloud apparitions, city changes, etc). A user may be interested in retrieving similar events and thus, may want to know when and where similar spatio-temporal patterns occurred. Moreover, as sub-graphs $\mathscr{G}_k$ contained in $\mathscr{G}$ are stochastic models for these spatio-temporal patterns, they can also possess a user semantic. An example of the connection between spatio-temporal patterns, user semantics and graphs is provided in Fig. 1. Therefore, we are interested in learning a semantic from a user in order to achieve a semantic labeling of sub-graphs representing spatio-temporal patterns which enables the recognition and the probabilistic retrieval of similar events.

The paper outline is as follows. After a short summary on the basic concepts of Bayesian inference, the article describes how the user semantics are modeled in a Bayesian context. As the semantic modeling is based on user-specific sub-graph similarities, a parametric distance model is first derived. Then, based on a Dirichlet model, a Bayesian interactive learning procedure estimates the similarity measurement parameters using user positive and negative sub-graphs examples. Therefore, the semantic labeling of sub-graphs can be achieved. Finally, after a section on practical applications, a short summary concludes the discussion.

Basic concepts of Bayesian Inference

In the Bayesian formalism, the concept of probability as the frequency of an event is extended to the concept of probability as the degree of certainty of a particular hypothesis H. The basic expressions are statements about conditional probabilities, e.g., $p(H \mid D)$, which specify the belief in the hypothesis H under the assumption that the data D is known. The heart of Bayesian techniques lies on the well known Bayes rule

$$p(H \mid D) = \frac{p(D \mid H)p(H)}{p(D)} \tag{1}$$

which relates the posterior $p(H \mid D)$ to the likelihood $p(D \mid H)$. Here, the hypothesis prior $p(H)$ reflects the prior belief in H and the data prior $p(D)$ acts as a normalizing constant. The power of Bayesian techniques in data (image time-series) processing comes from the fact that strong stochastic models are available for the data (image time-series) for low level modeling (e.g. Gaussian mixture models for feature space modeling) or higher level modeling (e.g. multinomial distribution with a Dirichlet conjugate prior for semantic modeling). They can be applied as models for the likelihood $p(D \mid H)$ for direct information extraction or for posterior $p(H \mid D)$ to incorporate a prior knowledge $p(H)$.

BAYESIAN INTERACTIVE LEARNING ON GRAPHS

The inference of the graph $\mathscr{G}$ is a robust and unsupervised coding of the SITS. Based on this objective signal characterization, we focus now on modeling interactively user interest for spatio-temporal patterns in the SITS. This framework is similar to the one adopted in the I2M system [3].

User-specific semantic modeling

In order to define a model for user semantic, a parametric similarity measure $S_\Phi(\mathscr{G}_0, \mathscr{G}_k)$ between two sub-graphs $\mathscr{G}_0$ and $\mathscr{G}_k$ is employed. This measure is an extension of the inexact matching algorithm proposed in [1]. The parameter vector Φ weights the contribution of each type of sub-graph features. A given parameter vector corresponds to a particular similarity formalizing a user conjecture.

By defining interactively a similarity, we will see in the two next sections that it is possible to link the subjective elements $\mathscr{A}_v$ representing a user conjecture to the objective sub-graph features $\mathscr{G}_k$ by learning the likelihood probabilities $p(\mathscr{G}_k \mid \mathscr{A}_v, M)$. For notation simplification, the conditioning of the likelihood by a model M is omitted in the following.

Based on these likelihood probabilities, using a Bayesian context enables the estimation of posterior probabilities $p(\mathscr{A}_v \mid \mathscr{G}_k)$ and thus, allows a semantic representation of the SITS content. Indeed, considering that a user provides positive and negative examples, corresponding to a positive $\mathscr{A}_v$ and a negative $\neg\mathscr{A}_v$ semantic, two likelihood probabilities $p(\mathscr{G}_k \mid \mathscr{A}_v)$ and $p(\mathscr{G}_k \mid \neg\mathscr{A}_v)$ can be derived for each sub-graphs. Moreover, graph priors can be obtained using the formula $p(\mathscr{G}_k) = \sum_v p(\mathscr{G}_k \mid \mathscr{A}_v)p(\mathscr{A}_v)$, where the summation is done over the positive and negative semantics. Thus, assuming a uniform prior

on the semantics, the posterior probabilities of the positive semantic are inferred using Bayes rule :

$$p(\mathscr{A}_v \mid \mathscr{G}_k) = \frac{p(\mathscr{G}_k \mid \mathscr{A}_v)\, p(\mathscr{A}_v)}{p(\mathscr{G}_k)} = \frac{p(\mathscr{G}_k \mid \mathscr{A}_v)}{p(\mathscr{G}_k \mid \mathscr{A}_v) + p(\mathscr{G}_k \mid \neg \mathscr{A}_v)}. \tag{2}$$

Thus, to achieve the posterior estimation, we need to define : (1) a parametric distance model $S_\Phi(\mathscr{G}_0, \mathscr{G}_k)$ for sub-graph similarity, (2) the interactive Bayesian learning of the likelihoods $p(\mathscr{G}_k \mid \mathscr{A}_v)$ and $p(\mathscr{G}_k \mid \neg \mathscr{A}_v)$ using the similarity function.

Parametric distance model for sub-graph similarity

The idea of inexact graph matching is to transform one of the sub-graphs into the other one by assigning a cost to each vertex or edge addition/removal. However, sub-graphs $\mathscr{G}_k$ are specific multidimensional temporal features which characterize parts of the dynamic MT cluster trajectories. The information is condensed in vertices and edges. A vertex is representing a TL cluster related to a given MT cluster. It is characterized by a pixel weight π, Gaussian parameters $\xi = (M, A)$ and a divergence measurement. Moreover, spatial information is contained in the corresponding TL classes. An edge, representing the evolution of the MT cluster between two image samples, is characterized by a time sampling delay T, a pixel flow γ , Gaussian parameter evolution $\delta(\xi)$ and intra-cluster changes MI. Thus, the inexact graph matching algorithm is extended to a parametric distance model between sub-graphs, weighting the different attribute contributions.
Denoting by $v_1 = \{v_i^1\}$ and $v_2 = \{v_i^2\}$ the vertex sets of sub-graphs $\mathscr{G}_1$ and $\mathscr{G}_2$, and denoting an extra set of vertices by $\lambda = \{\lambda_i\}$, a mapping function $\mathscr{F} = \{f\}$ composed by a given combination of elementary mapping functions $f : v^1 \to v^{2\lambda} = v^2 \cup \lambda$ is defined. A cost $C_\Phi(f(v_i^1) = v_j^{2\lambda})$ is assigned to each elementary transformations. The cost function depends on the vector parameter $\Phi = \{\phi_k\}$ and is composed by a weighted sum of similarities between vertices v_i^1 and $v_j^{2\lambda}$ and related edges. Preserving previous notations, when considering only incoming edges related to a node v_i, flows and Gaussian parameter evolutions related to edges are denoted by γ_{v_i} and $\delta_{v_i}(\xi)$. The cost $C_\Phi(f(v_i^1) = v_j^{2\lambda})$ is equal to

$$\phi_1 \Delta(\pi_{v_i^1}, \pi_{v_j^{2\lambda}}) + \phi_2 \Delta(\xi_{v_i^1}, \xi_{v_j^{2\lambda}}) + \phi_3 \Delta(\gamma_{v_i^1}, \gamma_{v_j^{2\lambda}}) + \phi_4 \Delta(\delta_{v_i^1}(\xi), \delta_{v_j^{2\lambda}}(\xi)) \tag{3}$$

where $\Delta(.)$ represents a distance model which is either a difference for scalars or a similarity measure between PDFs such as Kullbach-Leibler divergence. Because, time sampling delay and mutual information are characterizing all edges of a given MT class in a given interval, it must be added in the cost function only once per time interval. Furthermore, when considering 2 MT cluster at a time, their similarity denoted by $\delta_{MT}(\xi)$ is evaluated using Gaussian parameter similarity and reported once per time sampling in the cost function. Thus, when satisfying given conditions, the terms $\phi_5 \Delta(MI) + \phi_6 \Delta(T) + \phi_7 \Delta(\delta_{MT}(\xi))$ are added to the previous cost. The sub-graphs similarity is then defined, for a given vector parameter Φ, by finding the less expensive elementary mapping function combination over all possible mapping functions:

$$S_\Phi(\mathscr{G}_1, \mathscr{G}_2) = \sum_l S_{\phi_l}(\mathscr{G}_1, \mathscr{G}_2) = \min_{\mathscr{F}} \Big(\sum_i C_\Phi(f(v_i^1) = v_j^{2\lambda}) \Big), \tag{4}$$

where $S_{\phi_l}(\mathscr{G}_1,\mathscr{G}_2)$ is the contribution related to parameter ϕ_l, to the cost function. In order to estimate the minima, an optimization procedure is performed searching a minimum cost path in a tree containing all possible mapping functions configurations. Because, of the combinatorial explosion of configurations, the tree is pruned during the search, with the potential drawback that the optimization leads to a local minima.

Bayesian interactive learning of similarity

Tuning correctly these parameters in order to define a users-specific distance model might not be obvious for a user. Therefore a supervised learning procedure is needed to estimate the parameter vector Φ, enabling the evaluation of semantic likelihoods $p(\mathscr{G}_k \mid \mathscr{A}_v)$ and $p(\mathscr{G}_k \mid \neg\mathscr{A}_v)$. We detail in the following, the obtaining of the positive semantic likelihood. The negative semantic likelihood is obtained in a similar framework.

The sub-graph similarity function takes its value on an interval $I_m = [0,1/m]$. For a sub-graph $\mathscr{G}_k$ and a particular reference sub-graph $\mathscr{G}_0$, the parameter ϕ_i is assumed proportional to the partial cost function $S_{\phi_i}(\mathscr{G}_0,\mathscr{G}_k)$ and thus, $\phi_i \propto -S_{\phi_i}(\mathscr{G}_0,\mathscr{G}_k)$. The continuous parameters ϕ_i are discretized into r bins so that a particular parameter vector ϕ_i, taking its value in $\{\phi_i^1,...,\phi_i^r\}$, is assumed to follow a multinomial distribution. Thus, considering the user semantic $\mathscr{A}_v$, the conditioned PDF is defined by

$$p_{\mathscr{G}_0}(\phi_i = \phi_i^k \mid \omega,\mathscr{A}_v) = p(\Lambda_m(S_{\phi_i}(\mathscr{G}_0,\mathscr{G}_k)) = \phi_i^k \mid \omega,\mathscr{A}_v) = \omega_k, k = 1,...,r \tag{5}$$

where $\omega = \{\omega_2,...,\omega_r\}$ are the parameters and $\Lambda_m(.)$ is an operator, related to the interval value I_m, discretizing the partial cost function domain into r bins. For notation simplifications, $p_{\mathscr{G}_0}(\phi_i = \phi_i^k \mid \omega,\mathscr{A}_v)$ will be noted $p(\phi_i^k \mid \omega,\mathscr{A}_v)$. Note that the probabilities still depends on $\mathscr{G}_0$. Furthermore, statistical independence is assumed on the parameter conditioned distribution and thus, $p(\Phi \mid \mathscr{A}_v) = p(\Phi_i \mid \mathscr{A}_v)p(\Phi_j \mid \mathscr{A}_v)...$ so that the estimation of the joint probability distribution function is not needed.

We now move the discussion from assessing the probability of the parameter Φ_i to assessing the probability of the parameter ω. The estimation is learned by the training of the system by a user. A Bayesian framework is adopted because of its robustness when very limited user examples are available. The user provide a training dataset T of sub-graphs examples in accordance to his semantic. By decomposing the global cost of the similarity function into partial costs (Eq. 4), with the user sub-graphs examples, we define, for each parameters ϕ_i, a vector $N = \{N_1,...,N_r\}$ with N_k being the number of instance of ϕ_i^k.

The learning of a multinomial distribution (Eq. 5) uses for initialization, the simple conjugate prior distribution defined by a Dirichlet function with all hyper-parameters α_k equal to one which represents an uniform PDF.

The prior Dirichlet function is $p(\omega) = Dir(\omega \mid \alpha_1^{(0)},...,\alpha_r^{(0)}); \forall k \in [1,r], \alpha_k^{(0)} = 1$. After observing the instances $\{N_k^{(1)}\}$ in a training dataset $T^{(1)}$, according to Bayes rule, the posterior probability is

$$p(\omega \mid T^{(1)}) = \frac{p(T^{(1)} \mid \omega)p(\omega)}{p(T^{(1)})} = Dir(\omega \mid \alpha_1^{(0)} + N_1^{(1)},...,\alpha_r^{(0)} + N_r^{(1)}) \tag{6}$$

After observing another training dataset $T^{(2)}$, which is assumed to be independent from $T^{(1)}$ we obtain the new posterior

$$p(\omega \mid T^{(2)}, T^{(1)}) = \frac{p(T^{(2)} \mid \omega, T^{(1)}) p(\omega \mid T^{(1)})}{p(T^{(2)})} = Dir(\omega \mid \alpha_1^{(1)} + N_1^{(2)}, \ldots, \alpha_r^{(1)} + N_r^{(2)}) \quad (7)$$

where the new hyper-parameters were calculated by adding the number of times ϕ_i^k occurred in the training data set $T^{(2)}$. Therefore, each observed set of data $T^{(i)}$ can be incorporated as an update of the hyper-parameters : $\alpha_k^{(i)} = \alpha_k^{(i-1)} + N_k^{(i)}$.
Considering some training T with the associated hyper-parameter vector α, the estimation of $p(\phi_i^k \mid \mathscr{A}_v, T)$ is achieved using the minimum mean square error (MMSE) estimator of the parameter ω_k :

$$p(\phi_i^k \mid \mathscr{A}_v) = E[\omega_k] = \int \omega_k p(\omega \mid T) d\omega = \frac{\alpha_k}{\sum_k \alpha_k}. \quad (8)$$

Finally, by using the independence assumption, we obtain $p(\Phi \mid \mathscr{A}_v)$ by making the product of the $p(\phi_i^k \mid \mathscr{A}_v)$. After some training T, one can use the MMSE estimator to evaluate the parameter vector Φ of the similarity function. It is defined by $\widehat{\Phi}_{MMSE} = E\{p(\Phi \mid \mathscr{A}_v)\}$, where $E\{.\}$ is the expectation operator. Using this new estimation, a new similarity function $S_{\widehat{\Phi}_{MMSE}}(\mathscr{G}_0, \mathscr{G}_k)$ can be computed. Therefore, a semantic likelihood probability can be assigned to each sub-graph $\mathscr{G}_k$ according to

$$p(\mathscr{G}_k \mid \mathscr{A}_v) = \Upsilon\left(S_{\widehat{\Phi}_{MMSE}}(\mathscr{G}_0, \mathscr{G}_k)\right) \quad (9)$$

where $\Upsilon(.)$ is a linear operator inversely proportional, mapping the similarity function values into probabilities. Note that the probabilities are dependent of the reference graph $\mathscr{G}_0$. Changing of reference graph must be done regularly when the user adopts a browsing strategy for his search.
Therefore, sub-graphs can be retrieved according to the learned user-specific semantic posterior probabilities.

REASONING ON SITS

Experiments were performed on a temporal database composed of 38 superposable SPOT images of 2000x3000 pixels. They were obtained by daily acquisition and by filtering the snow-free and cloud-free images in a period of 286 days. The dynamic scene was acquired in a rural area near Bucharest (Romania). We considered a spatial subset of 200x200 pixels. Features are SPOT spectral reflectances.
First, the unsupervised part of the hierarchical modeling leaded to a graph characterization of SITS. Then, interactive learning was performed on sub-graphs, using a JAVA visual interface linking, sub-graphs to dynamic classifications and to the SITS. 3 different semantics were trained : cloud apparition, harvest and dephased harvest phenomena. The training of the cloud apparition semantic was performed using a time-window of

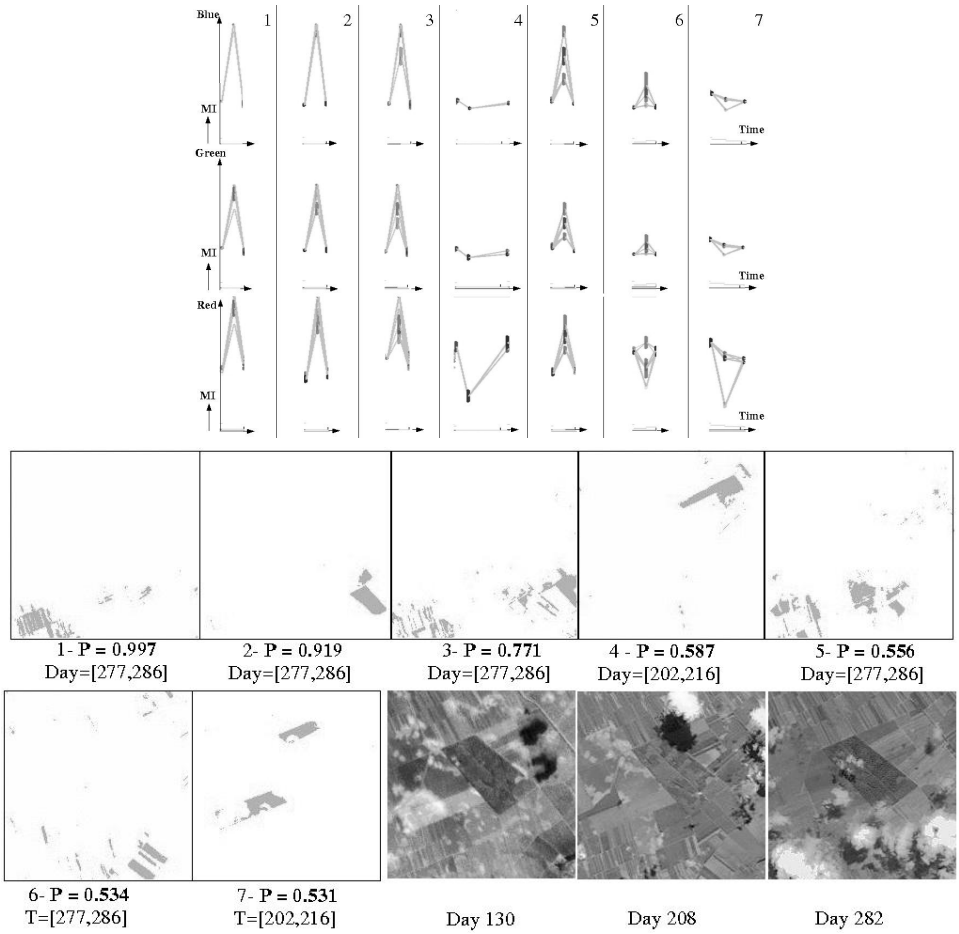

FIGURE 2. Training of a cloud apparition semantic. Above, the 7 most probable graph patterns; below, corresponding 7 most likely spatial classes with their probabilities and their time-windows, and retrieved image samples containing clouds.

3 samples where time delays were not considered (in the similarity function). Fig. 2 displays on the top, the retrieval of the 7 graph patterns of highest probabilities resulting from the search. Below, the corresponding MT spatial classes are presented together with their posterior probabilities, their time coordinates and the corresponding images where clouds appear. A second training was performed using an harvest semantic. On the top of Fig 3, the resulting 5 most probable spatial classes are presented together with their time-windows and their posterior probabilities. A typical graph pattern characterizing the phenomenon is also displayed. A time-window of 5 samples was used and time delays were not considered in the similarity function. The SITS between day 0 and day 125 where harvest occurred is displayed in Fig. 1. Almost all the retrieved harvests occurred in this period. A last training was performed for the retrieval of dephased harvests occurring in 2 different MT classes with similar cultures. A graph cycles characterizes this phenomena. On the right of Fig 3, the two sub-graphs of highest probability are presented with their spatial classes and their time coordinates corresponding to parts of the SITS displayed in Fig. 1.

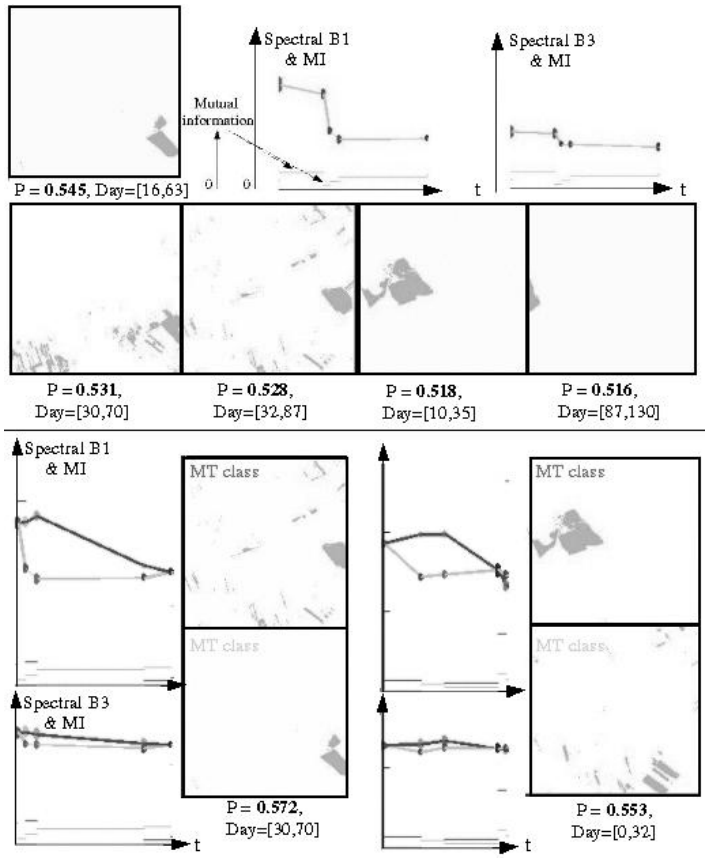

FIGURE 3. Harvest (above) and dephased harvest (below) semantic training

CONCLUSION

This work is an attempt to solve the complex problem of spatio-temporal reasoning on SITS. The proposed concept, developed in a Bayesian framework, models a user semantic by interactive learning on graphs. Based on the experiments, the method appears to be a fast and relevant way to retrieve user-specific spatio-temporal patterns[1].

REFERENCES

1. H. Bunke and G. Allerman, *Inexact graph matching for structural pattern recognition*, Pattern Recognition Lett. 1(4), pp. 245-253, 1983.
2. P. Héas, M. Datcu and A. Giros, *Trajectory of Dynamic Clusters in Image Time-Series*, In Proc. of MultiTemporal2003, Ispra, Italy, 2003.
3. M. Schroeder, H. Rehrauder, K. Seidel et M. Datcu, *Interactive learning and probabilistic retrieval in remote sensing image archives*, IEEE Trans. on GRS, Vol. 38, pp. 2288-2298, 2000.

[1] The authors would like to thank Alain Giros from the French Space Agency *(CNES)* for stimulating discussions and for carefully preprocessing the data.