

© 2012 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Title: Detection of Land-Cover Transitions in Multitemporal Remote Sensing Images with Active Learning Based Compound Classification

This paper appears in: IEEE Transactions on Geoscience and Remote Sensing

Date of Publication: 2012

Author(s): Begum Demir, Francesca Bovolo and Lorenzo Bruzzone

Volume:50, Issue: 5

Page(s): 1930 - 1941

DOI: 10.1109/TGRS.2011.2168534

# Detection of Land-Cover Transitions in Multitemporal Remote Sensing Images with Active Learning Based Compound Classification

Begüm DEMİR, *Member IEEE*, Francesca BOVOLO, *Member IEEE*,  
and Lorenzo BRUZZONE, *Fellow, IEEE*

Dept. of Information Engineering and Computer Science, University of Trento,  
Via Sommarive, 14, I-38123 Trento, Italy  
e-mail: demir@disi.unitn.it, francesca.bovolo@disi.unitn.it, lorenzo.bruzzone@ing.unitn.it.

**Abstract**— This paper presents a novel iterative active learning (AL) technique aimed at defining effective multitemporal training sets to be used for the supervised detection of land-cover transitions in a pair of remote sensing images acquired on the same area at different times. The proposed AL technique is developed in the framework of the Bayes rule for compound classification. At each iteration it selects the pair of spatially aligned unlabeled pixels in the two images that are classified with the maximum uncertainty. These pixels are then labeled by an external supervisor and included in the training set. Uncertainty of a pair of pixels is assessed by the joint entropy defined considering two possible different simplifying assumptions: i) class-conditional independence, and ii) temporal independence between multitemporal images. Accordingly, different algorithms are introduced. The proposed joint entropy based AL algorithms for compound classification are compared to each other and with a marginal entropy based AL technique (in which the entropy is computed separately on single-date images) applied to the post-classification comparison method. Experimental results obtained on two multispectral and multitemporal data sets show the effectiveness of the proposed technique.

***Index Terms*** –Multitemporal images, compound classification, active learning, joint entropy, change detection, remote sensing

## I. INTRODUCTION

The analysis of multitemporal remote sensing images to detect land-cover transitions plays an important role in many applications, such as assessment of damaged areas (*e.g.*, burned areas, flooded areas, *etc.*), analysis of urban expansion, study on shifting cultivation, *etc.* [1]. In the literature many algorithms that address the above-mentioned applications are available; generally they can be split into two categories: i) unsupervised algorithms, in which land-cover changes are detected by comparing the spectral reflectance values of remote sensing images [2]-[4], and ii) supervised algorithms, which require the availability of labeled training samples [5]-[8]. In the most of the cases unsupervised methods provide binary change-detection maps, where only the information about presence/absence of change is highlighted. On the contrary, supervised algorithms provide also the information about the kind of transitions occurred on the ground (*i.e.*, the labels of pixels before and after the change) at the cost of requiring reference data. The performance of unsupervised techniques is generally affected by the differences in atmospheric conditions, in sensor calibration, in ground conditions, *etc.* Supervised techniques are less affected by this kind of problems as they are based on the classifiers that recognize the transitions on the basis of the information present in the training set, which is used in the learning phase of the classification algorithm [5]-[8].

Only few approaches have been presented in the literature to detect the land-cover transitions by supervised techniques. The simplest supervised method to detect land-cover transitions (and also changes) is the post-classification comparison [5], which compares the classification maps obtained by independently classifying two remote-sensing images of the same area acquired at

different times. The accuracy obtained by this technique is close to the product of the accuracies obtained at the two times (if errors on the two classification maps are independent), as temporal dependence between multitemporal images is ignored. Temporal dependence between multitemporal images is exploited by the supervised direct multidata classification and compound classification techniques, which can obtain improved accuracy in the detection of land-cover transitions. The supervised direct multidata classification technique considers each transition as a class and trains a classifier to identify the transitions [5]. However, this technique has some limitations: i) the training samples at the two times should be related to the same points on the ground, and ii) the training samples should be statistically representative for all the possible transitions in the whole scene. In [6]-[8], the compound classification technique has been presented, which overcomes the constraints of direct multidata classification and increases the accuracy of the post-classification comparison technique (which is affected by the error-propagation phenomenon) by taking into account the temporal dependence between the images. In compound classification, temporal dependence is modeled by the prior joint probabilities of land-cover transitions between multitemporal images [6]-[8].

In this paper, we focus our attention on compound classification due to its effectiveness for the detection of land-cover transitions. One of the main motivations for which we consider the supervised compound classification technique is the increased interest that we expect in the future for supervised change-detection methods in remote sensing. This mainly depends on the properties of the last generation of passive sensors that can acquire images with either very high geometrical resolution (VHR) or very high spectral resolution (hyperspectral). VHR images are currently widely available at commercial level (*i.e.*, Quickbird, SPOT-5, Eros, *etc.*) whereas hyperspectral images will be available on a larger scale soon, as several satellite missions are under development. On the one hand, the many critical factors that affect the acquisition of VHR multitemporal images (*e.g.*, sensor view angle, sunrays angle) make it difficult to compare them by

completely unsupervised algorithms that work only on the radiance of the images without extracting the semantic meaning of the spectral differences (see [9] for greater details). On the other hand, the spectral signature measured by hyperspectral sensors is very detailed and many differences can be observed in the radiometric behavior of the images (*e.g.*, due to soil moisture differences in the ground conditions) even when no changes occurred on the ground. Thus, even in this case it is difficult to obtain reliable change-detection maps without supervised algorithms and thus reference data. Although many studies are in progress for defining unsupervised change-detection algorithms for the aforementioned kinds of data, the relevance of supervised approaches is expected to increase with respect to the past.

Although the compound classification technique is less critical with regard to the definition of the training set when compared to the multirate direct classification [6], the amount and quality of the available training samples are important to obtain accurate transition maps. However, the collection of multitemporal labeled samples is time consuming and costly; thus, to acquire a sufficient number of labeled training samples for each single-date image is a difficult task in real applications. To deal with this problem, active learning (AL) [10]-[27] and semi-supervised learning (SSL) [27]-[30] methods have been presented in the literature in the context of classification of single-date images. AL methods automatically select the most informative unlabeled samples to be manually labeled by a human expert in order to properly build up a non-redundant and effective training set, whereas SSL methods exploit both labeled and unlabeled samples in the learning of the classification algorithm [27]-[30]. In [27], the performances of the SSL and AL classification approaches have been compared, and it has been shown that SSL provides good results when the two analyzed images have similar properties. However, only AL is reliable when the spectral differences between the two images are significant. For this reason, in this paper we focus our attention on AL by generalizing its use to the context of compound classification. In greater detail, we present a novel AL technique for compound classification that

can be used to detect land-cover transitions. The proposed AL technique models the uncertainty of the labeling of pair of pixels in images acquired on the same area at different times by defining a joint entropy measure. The joint entropy uncertainty measure is defined in general and analyzed in two cases under different simplifying assumptions: i) the class-conditional independence in the time domain, and ii) the temporal independence. The experiments conducted on two different multitemporal and multispectral data sets show the effectiveness of proposed AL technique.

The paper is organized into seven sections. Section II gives background on AL, whereas Section III formulates the considered problem. The proposed AL method defined for compound classification is introduced in Section IV. Section V presents the description of the considered data sets and the design of experiments. Section VI illustrates the experimental results. Finally, Section VII draws the conclusion of this work.

## II. BACKGROUND ON ACTIVE LEARNING

In this section, we give the general definition for AL, and review some AL techniques presented in the literature for classification of single remote sensing images. AL techniques iteratively expand the size of an initial labeled training set  $T$  selecting the most informative samples from a pool  $U$  of unlabeled samples for manual labeling. At each iteration, the most informative unlabeled samples (for a given classifier  $G$ ) are selected based on a query function  $Q$ , labeled by a supervisor  $S$  and added to the current training set  $T$ . Finally, the supervised classifier  $G$  is retrained with the samples moved from  $U$  to  $T$ . It is worth nothing that the initial training set  $T$  requires few labeled samples for the first training of the classifier  $G$  and then is enriched iteratively by including the most informative samples selected from  $U$ . At the convergence, the training set  $T$  is made up of a minimum number of samples “optimal” for the considered classifier  $G$ . When the AL process is completed, the classifier  $G$  is trained once again and the classification of the image under investigation is carried out. The general flowchart of the AL-based classification approach is given in Fig. 1. The selection of the most informative samples from a pool  $U$  to be included in the

training set  $T$  on the basis of AL offers three main advantages: i) the labeling cost is reduced due to the avoidance of redundant samples, ii) the computational complexity of the learning phase is reduced due to the selection of an optimal subset of training samples (*i.e.*, a set with a small number of most representative samples), and iii) accurate classification accuracy can be obtained due to the improved class models estimated on a high quality training set on the basis of the classification rule used from the considered classifier. The supervisor  $S$  is usually a human expert who gives the true class labels to the selected samples. For remote sensing (RS) classification problems, the labeling of both the initial training set and of queried samples can be obtained by: 1) *in situ* ground surveys, 2) image photointerpretation, or 3) hybrid solutions (both photointerpretation and ground surveys).

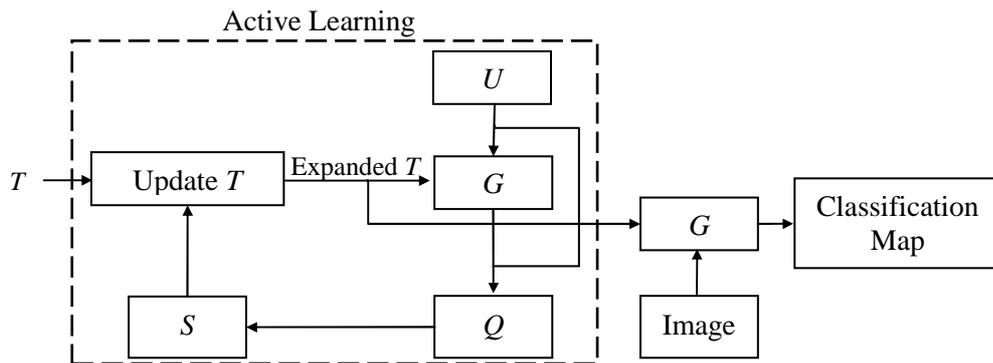


Fig. 1. General scheme of an AL based classification approach ( $T$ : Training set,  $G$ : Supervised classifier;  $Q$ : Query function;  $S$ :Supervisor;  $U$ : pool of unlabeled samples).

Active learning approaches can be divided into two main categories: i) uncertainty based approaches, and ii) query by committee based approaches. Uncertainty based approaches select the unlabeled samples that have the lowest confidence (*i.e.*, the maximal uncertainty) to be correctly classified by a given classifier (therefore most likely to be misclassified), and differ from each other for the adopted query functions. Uncertainty can be defined in different ways depending on the considered classifier. In [10] this kind of approaches has been implemented on the basis of the class-conditional posterior probabilities in the context of a maximum a-posteriori Bayesian classifier. The samples that do not show a predominant value of the estimated class-conditional

probability of one class over the others are selected as uncertain. The margin sampling (MS) technique has been proposed in [11], [12] for support vector machine (SVM) classifiers. It selects the unlabeled sample that is closest to the classification boundary. In [13] and [14] the most informative samples are selected among the most uncertain ones (which are the closest to the SVM classification boundary) on the basis of the standard  $k$ -means clustering technique and the angles between the induced classification hyperplanes, respectively. The marginal entropy based uncertainty criterion has been defined on the basis of conditional posterior probabilities of classes in [15],[16]. In these works, unlabeled samples that have the maximum entropy, *i.e.*, those having the maximum uncertainty among classes, are added to the training set at each iteration of AL. Query by committee based techniques select unlabeled samples that have the maximum disagreement within a committee of classifiers [17]-[18]. The disagreement among ensemble of classifiers is measured with Kullback–Leibler (KL) divergence in [17] and with entropy in [18].

The MS query function is extended to multiclass classification problems for multispectral images in [19] by selecting the most uncertain sample from each binary SVM. An AL method based on mutual information has been presented in [20] for the detection of unexploded ordnance. The AL technique proposed in [21] chooses the unlabeled sample that maximizes the information gain measured by the KL divergence. The KL divergence is calculated between the posterior probability distribution of the current training set and the training set obtained by including each unlabeled sample, one by one, into the training set. Two different AL techniques for multiclass RS classification problems are presented in [22]. In the first technique, the unlabeled samples that both have the smallest distance to the decision hyperplane of each binary SVM and do not share the same closest support vector are selected as uncertain and added to the training set. The second technique assesses the uncertainty on decisions of a committee of classifiers, *i.e.*, uncertain samples are those having maximum disagreement between a committee of classifiers. Disagreement among the classifiers is measured by the entropy in the distribution of the labels

provided by the committee members for each sample. Label acquisition costs sensitive AL techniques, which pay attention to the requirements of physical access to spatial locations for labeling process, have been proposed in [23],[24]. An AL technique based on a kernel-clustering has been presented in [25] to select the most informative representative samples among the most uncertain patterns. In this technique, the kernel-clustering is applied to the most uncertain samples selected according to the Multiclass-Level Uncertainty strategy [25]; then the most uncertain sample of each cluster is added to the training set. The AL methods presented in [13] and [14] have been modified in [25] to handle their limitations on real RS problems. A cluster-assumption based fast and reliable AL method defined on the basis of an histogram-thresholding algorithm has been presented in [26] for addressing critical problems where significantly biased initial training sets are available.

### III. PROBLEM FORMULATION

Differently from the AL techniques proposed in the RS literature that are devoted to single-date image classification, this paper aims to re-define AL in the context of the classification of multitemporal images. Let  $I_1 = \{x_{1,1}, x_{1,2}, \dots, x_{1,B}\}$  and  $I_2 = \{x_{2,1}, x_{2,2}, \dots, x_{2,B}\}$  denote two co-registered remote-sensing images made up of  $B$  pixels and acquired on the same area at two different times  $t_1$  and  $t_2$ , respectively. Let  $(x_{1,j}, x_{2,j})$  be the  $j$ -th pair of temporally correlated pixels made up of a pixel  $x_{1,j}$  acquired at time  $t_1$  and a spatially corresponding pixel  $x_{2,j}$  acquired at time  $t_2$ . Let  $\Omega = \{\omega_1, \omega_2, \dots, \omega_M\}$  be the set of possible land-cover classes at time  $t_1$ , and  $N = \{v_1, v_2, \dots, v_N\}$  be the set of possible land-cover classes at time  $t_2$ . Land-cover transitions (*i.e.*, changes in the labels) are observed if the two classes  $\omega_m$  ( $m = 1, \dots, M$ ) and  $v_n$  ( $n = 1, \dots, N$ ), to which the pair  $(x_{1,j}, x_{2,j})$  is assigned, are different. Here, differently from AL approaches for single-date image classification, the training set  $T$  and the pool  $U$  include pairs of pixels, and at

each iteration of the AL process, the training set  $T$  should be enriched by selecting the most informative pair of samples from the pool  $U$  of unlabeled pair of samples for manually labeling.

A basic trivial approach to apply AL to multitemporal image classification for the detection of land-cover transitions is to analyze the problem in the context of the post-classification comparison technique. Post-classification comparison is based on independent classification of each image. AL can be implemented by selecting the most uncertain samples from each single-date image by exploiting any AL technique proposed in the literature. In this way the training set of each image is independently enriched from the others. Although this method is simple, it has all disadvantages of the post-classification comparison technique explained in the previous section. For this reason, we focus on the definition of AL in the context of the Bayesian decision rule for compound classification. In particular, we present a novel technique to select the most uncertain pairs of pixels at each iteration of the AL process, which takes advantages of temporal dependence between images.

#### **IV. PROPOSED JOINT ENTROPY BASED ACTIVE LEARNING METHOD FOR COMPOUND CLASSIFICATION**

The proposed AL technique evaluates uncertainty according to the well-known entropy concept, which has been previously exploited in the literature in the context of AL-based single-date image classification [15]. In order to deal with multitemporal image classification, we propose a novel joint entropy based AL technique to measure the uncertainty. The general definition of joint entropy is firstly introduced. Then, different formulations defined under different simplifying assumptions on the temporal dependence between images are given. Before explaining the proposed technique, we briefly recall the compound classification technique for multitemporal images.

### A. Bayesian Decision Rule for Compound Classification

The Bayesian decision rule for compound classification identifies the best (in terms of Bayesian decision theory) pair of labels (classes) to be assigned to each pair of pixels  $(x_{1,j}, x_{2,j})$  by explicitly considering temporal dependence [6],[7], *i.e.*,

$$(x_{1,j}, x_{2,j}) \in (\omega_m, v_n) \text{ if } (\omega_m, v_n) = \arg \max_{\omega_i \in \Omega, v_k \in N} \{P(\omega_i, v_k | x_{1,j}, x_{2,j})\} \quad (1)$$

where  $P(\omega_i, v_k | x_{1,j}, x_{2,j})$  is the joint conditional posterior probability of the pair of classes  $(\omega_i, v_k)$ , given the pair of pixels  $(x_{1,j}, x_{2,j})$ . In the context of automatic detection of land-cover transitions, the estimation of the statistical quantities involved in (1) is a complex task due to the difficulty in collecting enough training samples for properly modeling the multitemporal dependence between all possible temporal combinations of classes. Therefore, according to the literature [6],[7], we adopt the conventional assumption of class-conditional independence in the time domain to simplify the estimation of the joint conditional posterior probabilities. Under this assumption, (1) can be rewritten as [6],[7]

$$\begin{aligned} & (x_{1,j}, x_{2,j}) \in (\omega_m, v_n) \text{ so that } (\omega_m, v_n) \\ & = \arg \max_{\omega_i \in \Omega, v_k \in N} \left\{ \frac{p(x_{1,j} | \omega_i) p(x_{2,j} | v_k) P(\omega_i, v_k)}{\sum_{\omega_s \in \Omega} \sum_{v_r \in N} [p(x_{1,j} | \omega_s) p(x_{2,j} | v_r) P(\omega_s, v_r)]} \right\} = \arg \max_{\omega_i \in \Omega, v_k \in N} \{p(x_{1,j} | \omega_i) p(x_{2,j} | v_k) P(\omega_i, v_k)\} \end{aligned} \quad (2)$$

where  $p(x_{1,j} | \omega_i)$  and  $p(x_{2,j} | v_k)$  are the single-date class-conditional density functions, and  $P(\omega_i, v_k)$  is the joint prior probability of having classes  $\omega_i$  at  $t_1$  and  $v_k$  at  $t_2$  in the multitemporal images. Joint prior probabilities of land-cover transitions can be estimated from the images under investigation by an iterative procedure defined on the basis of the expectation-maximization (EM) algorithm. The recursive equation to estimate  $P(\omega_i, v_k)$  is defined as [7]

$$P_z(\omega_i, v_k) = \frac{1}{B} \sum_{j=1}^B \frac{p(x_{1,j}|\omega_i)p(x_{2,j}|v_k)P_{z-1}(\omega_i, v_k)}{\sum_{\omega_s \in \Omega} \sum_{v_r \in N} [p(x_{1,j}|\omega_s)p(x_{2,j}|v_r)P_{z-1}(\omega_s, v_r)]} \quad (3)$$

where  $P_z(\omega_i, v_k)$  and  $P_{z-1}(\omega_i, v_k)$  are the joint prior probability estimates at the  $z$ th and  $(z-1)$ th iterations, respectively. The estimates are initialized by assigning equal joint prior probabilities to each pair of classes (eventually including possible specific constraints on some transitions according to prior information). Then (3) is iterated until convergence, which is achieved when the maximum difference between the estimates at two successive iterations is below a given threshold [7]. It is worth nothing that it is unfeasible to estimate joint prior probabilities directly from the multitemporal training set due to the difficulty in having training sets with a sufficient number of representative samples for each possible combination of classes [7]. For detailed information on compound classification, the reader is referred to [6],[7].

### *B. Joint Entropy based AL Method for Compound Classification*

The proposed AL method is based on the selection of unlabeled pairs of samples that have maximum uncertainty on their labels assigned by the Bayesian rule for compound classification. In order to consider temporal dependence in modeling the uncertainty of samples, the method is defined by using the joint entropy  $H(x_{1,j}, x_{2,j})$  of the decisions obtained by compound classification for the generic pair of pixels in corresponding positions  $(x_{1,j}, x_{2,j})$ . For each pair of pixels  $(x_{1,j}, x_{2,j})$  in the multitemporal images, the joint entropy is defined as

$$H(x_{1,j}, x_{2,j}) = - \sum_{\omega_i \in \Omega} \sum_{v_k \in N} P(\omega_i, v_k | x_{1,j}, x_{2,j}) \log P(\omega_i, v_k | x_{1,j}, x_{2,j}) \quad (4)$$

If  $H(x_{1,j}, x_{2,j})$  is small, the corresponding pair of pixels is classified with high confidence, *i.e.*, the decision on compound classification of these samples is reliable. If  $H(x_{1,j}, x_{2,j})$  is high, the decision is not reliable, and therefore the corresponding pair of pixels is considered uncertain and

critical for the classifier. Samples that satisfy the latter condition are suitable to be included in the training set after manual labeling by an external supervisor. In the proposed AL method, at each iteration the parameters (the class probability density functions and the joint prior probability of classes) are estimated for each pair of pixels  $(x_{1,j}, x_{2,j}) \in U$  (where  $U$  can be associated with the pixels of the entire image or with a subset of them). Then the joint entropies of all the pairs are calculated. The most uncertain  $h$  unlabeled pairs of samples, *i.e.*, those that have the maximum joint entropy, are extracted and given to the supervisor  $S$  for labeling. This process is iterated until convergence, which is reached when either i) the values of class parameters do not change anymore, or ii) the desired number of samples is labeled (*i.e.*, the upper bound of the cost for labeling samples is achieved). Fig. 2 shows the architecture of the proposed joint entropy based AL method.

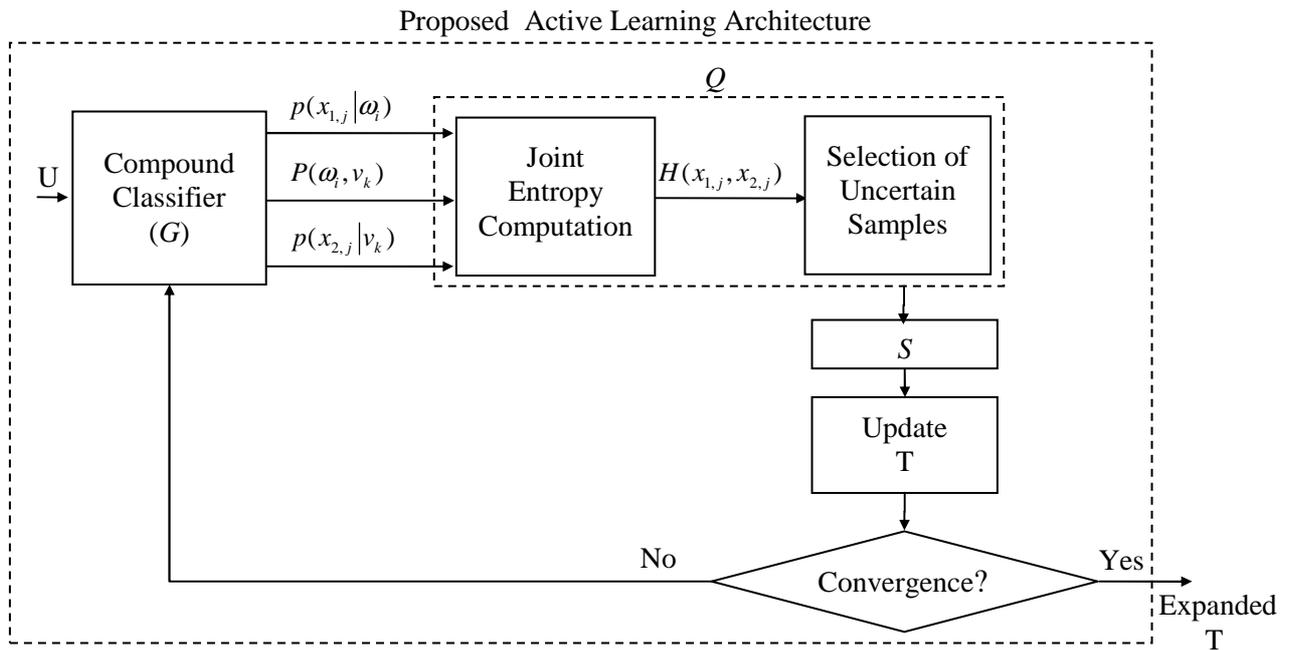


Fig. 2. Architecture of the proposed joint entropy based AL method

The definition in (4) depends on the joint conditional posterior probabilities of the pairs of classes, which play a key role in the definition of joint entropy. As mentioned above, in general it is difficult to have a sufficient number of labeled samples for a reliable estimation of these

quantities. Accordingly, we adopted simplifying assumptions that result in different algorithms for the proposed AL technique. These algorithms are defined under either i) the assumption of class-conditional independence, or ii) the assumption of temporal independence between multitemporal images.

*AL Algorithm Defined Under the Assumption of Class-Conditional Independence (JEAL):*

According to the simplified approach usually adopted in compound classification, initially we work under the assumption of class-conditional independence in the time domain, *i.e.*,  $p(x_{1,j}, x_{2,j} | \omega_i, v_k) = p(x_{1,j} | \omega_i) p(x_{2,j} | v_k)$ . Under this assumption, we can write:

$$P(\omega_i, v_k | x_{1,j}, x_{2,j}) = \frac{p(x_{1,j} | \omega_i) p(x_{2,j} | v_k) P(\omega_i, v_k)}{\sum_{\omega_s \in \Omega} \sum_{v_r \in N} [p(x_{1,j} | \omega_s) p(x_{2,j} | v_r) P(\omega_s, v_r)]} \quad (5)$$

In this condition  $P(\omega_i, v_k)$  is the only term that models the temporal dependence between the two images, *i.e.*, we use only the prior knowledge on the possible pair of labels and not the temporal dependence in the measured reflectance. Thus (4) can be rewritten as

$$H(x_{1,j}, x_{2,j}) = - \sum_{\omega_i \in \Omega} \sum_{v_k \in N} \left\{ \frac{p(x_{1,j} | \omega_i) p(x_{2,j} | v_k) P(\omega_i, v_k)}{\sum_{\omega_s \in \Omega} \sum_{v_r \in N} [p(x_{1,j} | \omega_s) p(x_{2,j} | v_r) P(\omega_s, v_r)]} \log \left( \frac{p(x_{1,j} | \omega_i) p(x_{2,j} | v_k) P(\omega_i, v_k)}{\sum_{\omega_s \in \Omega} \sum_{v_r \in N} [p(x_{1,j} | \omega_s) p(x_{2,j} | v_r) P(\omega_s, v_r)]} \right) \right\} \quad (6)$$

$H(x_{1,j}, x_{2,j})$  expresses the confidence of the decision on  $(x_{1,j}, x_{2,j})$  taking into account the temporal dependence between multitemporal images modeled by the joint prior probabilities  $P(\omega_i, v_k)$ . The temporal dependence between reflectance measures (*i.e.*, the possible correlation between the signals measured on the two pixels considered) is neglected. This is a reasonable tradeoff between the complexity of the estimation problem of joint conditional probabilities and the simplistic assumption of temporal independence.

*AL Algorithm Defined Under the Assumption of Temporal Independence Between Multitemporal Images (JEAL<sub>Ind</sub>):*

In order to further simplify the estimation of the joint conditional posterior probabilities given in (4), we can introduce a stronger simplifying assumption considering also the independence of a-priori class probabilities in the two images (*i.e.*,  $P(\omega_i, v_k) = P(\omega_i)P(v_k)$ ). According to this additional assumption, the term  $P(\omega_i, v_k | x_{1,j}, x_{2,j})$  can be rewritten as:

$$P(\omega_i, v_k | x_{1,j}, x_{2,j}) = \frac{p(x_{1,j} | \omega_i) p(x_{2,j} | v_k) P(\omega_i) P(v_k)}{\sum_{\omega_s \in \Omega, v_r \in N} [p(x_{1,j} | \omega_s) p(x_{2,j} | v_r) P(\omega_s) P(v_r)]} \quad (7)$$

Thus the joint entropy defined in (6) becomes:

$$H(x_{1,j}, x_{2,j}) = - \sum_{\omega_i \in \Omega} \sum_{v_k \in N} \left\{ \frac{p(x_{1,j} | \omega_i) p(x_{2,j} | v_k) P(\omega_i) P(v_k)}{\sum_{\omega_s \in \Omega} \sum_{v_r \in N} [p(x_{1,j} | \omega_s) p(x_{2,j} | v_r) P(\omega_s) P(v_r)]} \log \left( \frac{p(x_{1,j} | \omega_i) p(x_{2,j} | v_k) P(\omega_i) P(v_k)}{\sum_{\omega_s \in \Omega} \sum_{v_r \in N} [p(x_{1,j} | \omega_s) p(x_{2,j} | v_r) P(\omega_s) P(v_r)]} \right) \right\} \quad (8)$$

Equation (8) can be rewritten as:

$$H(x_{1,j}, x_{2,j}) = - \left\{ \sum_{\omega_i \in \Omega} \frac{p(x_{1,j} | \omega_i) P(\omega_i)}{\sum_{\omega_s \in \Omega} [p(x_{1,j} | \omega_s) P(\omega_s)]} \log \left( \frac{p(x_{1,j} | \omega_i) P(\omega_i)}{\sum_{\omega_s \in \Omega} [p(x_{1,j} | \omega_s) P(\omega_s)]} \right) \sum_{v_k \in N} \frac{p(x_{2,j} | v_k) P(v_k)}{\sum_{v_r \in N} [p(x_{2,j} | v_r) P(v_r)]} \right\} + \left\{ \sum_{v_k \in N} \frac{p(x_{2,j} | v_k) P(v_k)}{\sum_{v_r \in N} [p(x_{2,j} | v_r) P(v_r)]} \log \left( \frac{p(x_{2,j} | v_k) P(v_k)}{\sum_{v_r \in N} [p(x_{2,j} | v_r) P(v_r)]} \right) \sum_{\omega_i \in \Omega} \frac{p(x_{1,j} | \omega_i) P(\omega_i)}{\sum_{\omega_s \in \Omega} [p(x_{1,j} | \omega_s) P(\omega_s)]} \right\} \quad (9)$$

As expected, the last element of each term in (9) equals to 1. Thus, since marginal entropies

$H(x_{1,j})$  and  $H(x_{2,j})$  are defined as

$$\begin{aligned}
H(x_{1,j}) &= -\sum_{\omega_i \in \Omega} P(\omega_i | x_{1,j}) \log P(\omega_i | x_{1,j}) = -\sum_{\omega_i \in \Omega} \frac{p(x_{1,j} | \omega_i) P(\omega_i)}{\sum_{\omega_s \in \Omega} [p(x_{1,j} | \omega_s) P(\omega_s)]} \log \left( \frac{p(x_{1,j} | \omega_i) P(\omega_i)}{\sum_{\omega_s \in \Omega} [p(x_{1,j} | \omega_s) P(\omega_s)]} \right) \\
H(x_{2,j}) &= -\sum_{v_k \in N} P(v_k | x_{2,j}) \log P(v_k | x_{2,j}) = -\sum_{v_k \in N} \frac{p(x_{2,j} | v_k) P(v_k)}{\sum_{v_r \in N} [p(x_{2,j} | v_r) P(v_r)]} \log \left( \frac{p(x_{2,j} | v_k) P(v_k)}{\sum_{v_r \in N} [p(x_{2,j} | v_r) P(v_r)]} \right)
\end{aligned} \tag{10}$$

under the considered assumptions of independence the joint entropy is equal to the sum of marginal entropies computed on each image, *i.e.*,

$$H(x_{1,j}, x_{2,j}) = H(x_{1,j}) + H(x_{2,j}). \tag{11}$$

Under the above-mentioned assumptions, we derive a simplified version of the joint entropy based AL method (JEAL<sub>Ind</sub>), which selects the most uncertain pairs of samples according to the sum of marginal entropies, assuming temporally independence between multitemporal images. This means assuming that mutual information among them is zero. Indeed, from the information theory we can write [31]:

$$H(x_{1,j}, x_{2,j}) = H(x_{1,j}) + H(x_{2,j}) - MI(x_{1,j}, x_{2,j}) \tag{12}$$

where  $MI(x_{1,j}, x_{2,j})$  shows the mutual information between  $x_{1,j}$  and  $x_{2,j}$ . A comparison between (11) and (12) confirms that in the case of temporal independence assumption  $MI(x_{1,j}, x_{2,j}) = 0$ . As a result of the adopted assumptions, the parameter estimation step is simplified whereas the advantage of exploiting the prior information on the class temporal dependence is lost.

By analyzing JEAL and JEAL<sub>Ind</sub> in greater detail, we can have a better understanding of the difference between AL applied to single images or used in the context of compound classification. In the JEAL<sub>Ind</sub> algorithm, the sum of marginal entropies  $H(x_{1,j})$  and  $H(x_{2,j})$  models the uncertainty of a pair of spatially corresponding pixels  $(x_{1,j}, x_{2,j}) \in U$  and the samples with maximum sum of marginal entropies are selected as uncertain. High joint entropy values can be

obtained in the following cases: i) a sample has an uncertain label simultaneously for both single-date images [*i.e.*, marginal entropies  $H(x_{1,j})$  and  $H(x_{2,j})$  are both high], or ii) a sample shows uncertainty at least on one of the single-date images (*i.e.*, only one marginal entropy value is high, thus the sum of the marginal entropies can be high). If the joint entropy is small, the corresponding pair of samples is classified with high confidence on both images, *i.e.*, marginal entropies  $H(x_{1,j})$  and  $H(x_{2,j})$  are both low. On the contrary, in the case of JEAL the  $MI(x_{1,j}, x_{2,j})$  may be different from zero. This means that the final uncertainty on the selected pair of pixels depends not only on the marginal entropies but also on the mutual information which is modeled by  $P(\omega_t, v_k)$ . In this case  $MI(x_{1,j}, x_{2,j})$  is zero only if class labels are independent on the two images. We also observe that in the case in which  $P(\omega_t, v_k)$  have the same values for all possible pairs of classes (land-cover transitions with the same probability) the pairs of samples selected by JEAL and JEAL<sub>Ind</sub> are the same. Despite considering marginal entropies, JEAL<sub>Ind</sub> still exploits some information from multitemporal images, *i.e.*, it analyzes pixels with the same coordinate on the two images that correspond to the same areas on the ground. A further simplification can be obtained measuring the uncertainty by independently exploiting the marginal entropies on each considered image. In other words, we can also neglect the only remaining temporal information in JEAL<sub>Ind</sub>, which is the spatial correspondence of pixels. The marginal entropy based AL algorithm (MEAL), which selects the most uncertain unlabeled samples from a single-date image on the basis of the standard marginal entropy (*i.e.*, ignoring the temporal dependence between images) has already been used for single-date image classification in [15]. In order to use MEAL in compound classification, marginal entropy of each unlabeled sample can be calculated according to (11). At each iteration, the most uncertain  $h/2$  unlabeled samples of the image acquired at time  $t_1$  and the most uncertain  $h/2$  unlabeled samples of the image acquired at time  $t_2$  are selected. These samples are merged into a set of  $h$  uncertain pairs of samples. Since the same pairs can be selected more than one time,

the redundant (repetitive) ones are replaced with the next most uncertain pair. Finally, the supervisor  $S$  adds labels to the uncertain pairs of samples that are then added to the training set.

## V. DATA SET DESCRIPTION AND DESIGN OF EXPERIMENTS

### A. Data set description

Experimental analyses are conducted on two multitemporal data sets. The first data set is made up of two multispectral images acquired by the Quickbird multispectral sensor on the city of Trento, Italy, in October 2005 and July 2006. In the pre-processing phase both images were pan-sharpened and co-registered. The two images of this data set share six land-cover classes [*i.e.*, water, bare soil, vegetation, road, shadow, building]. Five different kinds of land-cover transitions can be identified between the two images (*i.e.*, from vegetation to bare soil, from bare soil to vegetation, from shadow to bare soil, from shadow to water, and from shadow to vegetation). This data set has a test set  $TS$  of 2083 pairs of samples and a pool  $U$  of 1868 pairs of samples. Table I and Table II show the number of samples corresponding to each land-cover transition occurred between the images for the pool and the test set, respectively. From the pool  $U$ , 10 pairs of samples related to each land-cover transition are randomly selected as initial training samples (therefore the training set  $T$  has 100 pairs of samples) and the remaining samples are considered as unlabeled. As one can see from the tables, the land-cover transition from shadow to vegetation is not represented in the pool  $U$ . Nonetheless, as it occurred, it has been included in the test set  $TS$  for accuracy assessment purposes.

The second data set is made up of two co-registered multispectral images acquired by the Landsat-5 satellite on the Island of Sardinia, Italy, in September 1995 and July 1996. The images share five land-cover classes (*i.e.*, pasture, forest, urban area, water, vineyard) and no land-cover changes are observed. Table III and Table IV show the land-cover classes and the related number of sample pairs used in the experiments for the pool and the test set, respectively. The test set has

1949 samples and the pool has 2249 samples. From the pool  $U$ , 20 pairs of samples related to each land-cover transition are randomly selected as initial training samples (therefore the training set has 100 pairs of samples) and the remaining samples are considered as unlabeled.

TABLE I. NUMBER OF SAMPLE PAIRS IN  $U$  FOR EACH LAND-COVER TRANSITION (TRENTO DATA SET).

<b>July 2006</b> <b>October 2005</b>	<b>Water</b> ( $v_1$ )	<b>Bare soil</b> ( $v_2$ )	<b>Vegetation</b> ( $v_3$ )	<b>Road</b> ( $v_4$ )	<b>Shadow</b> ( $v_5$ )	<b>Building</b> ( $v_6$ )	<b>Total</b>
<b>Water (<math>\omega_1</math>)</b>	352	-	-	-	-	-	352
<b>Bare soil (<math>\omega_2</math>)</b>	-	234	276	-	-	-	510
<b>Vegetation (<math>\omega_3</math>)</b>	-	62	276	-	-	-	338
<b>Road (<math>\omega_4</math>)</b>	-	-	-	241	-	-	241
<b>Shadow (<math>\omega_5</math>)</b>	77	70	-	-	101	-	248
<b>Building (<math>\omega_6</math>)</b>	-	-	-	-	-	179	179
<b>Total</b>	429	366	552	241	101	179	1868

TABLE II. NUMBER OF SAMPLE PAIRS IN  $TS$  FOR EACH LAND-COVER TRANSITION (TRENTO DATA SET).

<b>July 2006</b> <b>October 2005</b>	<b>Water</b> ( $v_1$ )	<b>Bare soil</b> ( $v_2$ )	<b>Vegetation</b> ( $v_3$ )	<b>Road</b> ( $v_4$ )	<b>Shadow</b> ( $v_5$ )	<b>Building</b> ( $v_6$ )	<b>Total</b>
<b>Water (<math>\omega_1</math>)</b>	466	-	-	-	-	-	466
<b>Bare soil (<math>\omega_2</math>)</b>	-	254	225	-	-	-	479
<b>Vegetation (<math>\omega_3</math>)</b>	-	213	161	-	-	-	374
<b>Road (<math>\omega_4</math>)</b>	-	-	-	223	-	-	223
<b>Shadow (<math>\omega_5</math>)</b>	108	50	52	-	93	-	303
<b>Building (<math>\omega_6</math>)</b>	-	-	-	-	-	238	238
<b>Total</b>	574	517	438	223	93	238	2083

TABLE III. NUMBER OF SAMPLE PAIRS IN  $U$  FOR EACH LAND-COVER TRANSITION (SARDINIA DATA SET).

<b>July 1996</b> <b>September 1995</b>	<b>Pasture</b> ( $v_1$ )	<b>Forest</b> ( $v_2$ )	<b>Urban Area</b> ( $v_3$ )	<b>Water Body</b> ( $v_4$ )	<b>Wineyard</b> ( $v_5$ )	<b>Total</b>
<b>Pasture (<math>\omega_1</math>)</b>	554	-	-	-	-	554
<b>Forest (<math>\omega_2</math>)</b>	-	304	-	-	-	304
<b>Urban Area (<math>\omega_3</math>)</b>	-	-	408	-	-	408
<b>Water Body (<math>\omega_4</math>)</b>	-	-	-	804	-	804
<b>Wineyard (<math>\omega_5</math>)</b>	-	-	-	-	179	179
<b>Total</b>	554	304	408	804	179	2249

TABLE IV. NUMBER OF SAMPLE PAIRS IN  $TS$  FOR EACH LAND-COVER TRANSITION (SARDINIA DATA SET).

<b>September 1995</b> \ <b>July 1996</b>	<b>Pasture</b> ( $v_1$ )	<b>Forest</b> ( $v_2$ )	<b>Urban Area</b> ( $v_3$ )	<b>Water Body</b> ( $v_4$ )	<b>Wineyard</b> ( $v_5$ )	<b>Total</b>
<b>Pasture</b> ( $\omega_1$ )	589	-	-	-	-	589
<b>Forest</b> ( $\omega_2$ )	-	274	-	-	-	274
<b>Urban Area</b> ( $\omega_3$ )	-	-	418	-	-	418
<b>Water Body</b> ( $\omega_4$ )	-	-	-	551	-	551
<b>Wineyard</b> ( $\omega_5$ )	-	-	-	-	117	117
<b>Total</b>	589	274	418	551	117	1949

### B. Design of experiments

We carried out experiments with batch size values  $h=10$  and  $h=20$ . The prior probabilities ( $P(\omega_i)$ ,  $\omega_i \in \Omega$ , and  $P(v_k)$ ,  $v_k \in N$ ) and the class probability density functions ( $p(x_{1,j}|\omega_i)$ ,  $p(x_{2,j}|v_k)$ ) are estimated from the training set. According to the remote sensing literature [32], class probability density functions are assumed to be Gaussian. The joint prior probabilities of classes  $P(\omega_i, v_k)$ ,  $\omega_i \in \Omega$ ,  $v_k \in N$  are calculated by exploiting the EM algorithm and the threshold value to stop iterations is fixed to 0.001 as in [7]. At each iteration of the AL process, the estimates of prior probabilities, class probability density functions and joint prior probabilities of classes are updated. All experimental results are given as the average accuracy obtained in 10 trials related to 10 initial randomly selected training sets. The size of final training set is fixed to 400 for both data sets.

For both Trento and Sardinia multitemporal data sets, we carried out two kinds of experiments: i) the first set of trials assesses the effectiveness of the proposed AL method defined under the considered simplifying assumptions, [*i.e.*, the JEAL algorithm (which uses the assumption of class-conditional independence in the time domain), the JEAL<sub>Ind</sub> algorithm (which assumes temporally independence between the images), and the MEAL algorithm (which also neglects spatial correspondence among pixels)], and ii) the second set of trials is devoted to the

comparison of the proposed AL method defined for compound classification with a marginal entropy based AL method applied to a post-classification comparison technique (AL-PCC). The post-classification comparison technique is defined by using a standard Bayesian maximum posterior probability classifier for each image under the Gaussian assumption for class distributions. AL-PCC is implemented by initially selecting the set of  $h$  uncertain samples from each single image (independently from the other image) on the basis of marginal entropy (see (10)), and then performing post-classification comparison.

In order to further assess the reliability of the proposed technique, we also compared the robustness to registration noise (*i.e.*, effects of non perfect alignment between multitemporal images) of both the proposed method and the AL-PCC technique. To this end, image  $I_1$  was shifted of a given number of pixels in the horizontal and the vertical directions. In the following, the value of the displacement will be indicated with a pair of numbers  $(n,m)$  that indicate the amount and the direction of the shift in pixels.  $n$  is positive for shifts in right direction and negative otherwise, while  $m$  is positive for upward shifts and negative otherwise.

## VI. EXPERIMENTAL RESULTS

### A. Trento Data Set

In the first set of trials, we compare the effectiveness of proposed AL algorithms with each other in the context of compound classification. Fig. 3 shows the average (on 10 trials) accuracies in the detection of land-cover transitions versus the number of training samples obtained by JEAL, JEAL<sub>Ind</sub> and MEAL algorithms when  $h=10$  and  $h=20$ . The results show that accuracies of JEAL, JEAL<sub>Ind</sub> and MEAL are quite similar at early iterations, whereas those of JEAL significantly increase with the size of the training set. Moreover, JEAL reaches convergence in a smaller number of iterations (*i.e.*, with a smaller number of labeled pairs of samples) than the other algorithms. MEAL algorithm provides in general much higher accuracies than JEAL<sub>Ind</sub> in case of

$h=10$ , and slightly higher accuracies in case of  $h=20$ . As an example (see Fig.3a that refers to  $h=10$ ), the JEAL algorithm provides an accuracy of 97.60% with only 190 samples, whereas JEAL<sub>IND</sub> and MEAL reach a similar accuracy with almost 300 samples. Moreover, the JEAL algorithm yields an accuracy of 98.22% with 400 samples, whereas the accuracies obtained by using the full pool as training set (1868 samples) is 98.51%. The higher accuracy obtained by the query based on MEAL than that of the query based on JEAL<sub>IND</sub> might appear unexpected. However it can be explained by the fact that if we do not model the dependence by the joint prior probabilities of classes, *i.e.*, if we neglect almost all the temporal information, working on the joint entropies of pixels may result misleading (and thus less effective) than selecting the most uncertain pixels for each image.

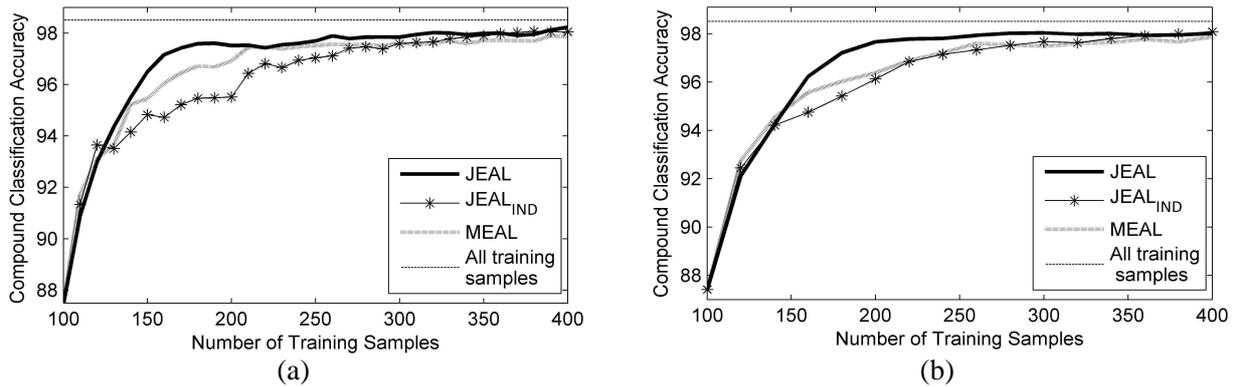


Fig. 3. Average (on 10 trials) overall accuracy in the detection of land-cover transitions obtained by using the JEAL, JEAL<sub>IND</sub> and MEAL algorithms when (a)  $h=10$  and (b)  $h=20$  and the full pool as training set (Trento data set).

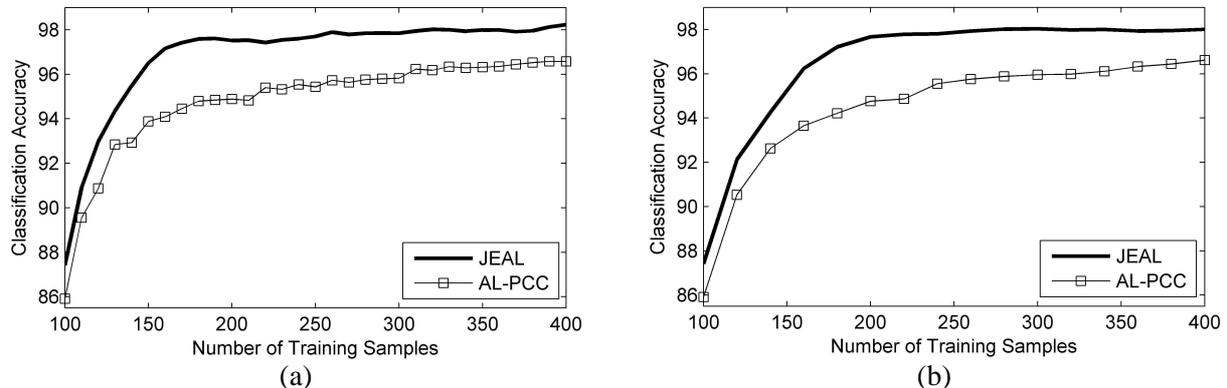


Fig. 4. Average (on 10 trials) overall accuracy in the detection of land-cover transitions obtained by the JEAL with compound classification (JEAL) and AL with post-classification comparison (AL-PCC) when (a)  $h=10$  and (b)  $h=20$  (Trento data set).

In the second set of trials, we compare the effectiveness of JEAL and AL-PCC. Fig. 4 shows the obtained results in terms of the average overall accuracy in the detection of the land-cover transitions. By analyzing the figure, one can observe that the JEAL algorithm leads to the highest accuracies for all the iterations and significantly outperforms the AL-PCC method. These results demonstrate: i) the effectiveness of compound classification technique, and ii) the effectiveness of joint entropy criterion for query in AL. Table V and Table VI show the average (on 10 trials) accuracies for each land-cover transitions obtained by the JEAL and AL-PCC, respectively at the 8th iteration (*i.e.*, the training data set includes 180 pairs of samples ). As one can see from the tables, the accuracies obtained by the JEAL algorithm are significantly higher than those yielded by AL-PCC for all the transitions but the shadow to water and the shadow to vegetation ones. This is due to the fact that the shadow class is detected with a very high accuracy (almost 100%) in the October image; thus the transitions that involve the shadow class result in a high accuracy even with the post-classification comparison technique. Nonetheless, the general effectiveness of the techniques should be evaluated taking into account the overall accuracy.

Table VII shows the performances of both JEAL and AL-PCC at the 8th iteration (*i.e.*, the training data set includes 180 pairs of samples) under various amounts of simulated misregistration, *i.e.*, with different values of  $(n,m)$ . From the table, one can observe that any possible misregistration between images decreases the performance of both methods when compared to the ones achieved with perfectly aligned images [*i.e.*,  $(n,m)=(0,0)$ ]. However, JEAL always performs better than AL-PCC, thus pointing out its robustness to the presence of registration noise. As an example, in the case of  $(n,m)=(-1,0)$ , the accuracy of JEAL is 96.20% whereas that of AL-PCC is 92.97%.

TABLE V. AVERAGE (ON 10 TRIALS) ACCURACIES (%) OF EACH LAND-COVER TRANSITION OBTAINED BY JEAL (TRENTO DATA SET)

<b>July 2006</b> <b>October 2005</b>	<b>Water</b> ( $v_1$ )	<b>Bare soil</b> ( $v_2$ )	<b>Vegetation</b> ( $v_3$ )	<b>Road</b> ( $v_4$ )	<b>Shadow</b> ( $v_5$ )	<b>Building</b> ( $v_6$ )
<b>Water</b> ( $\omega_1$ )	99.89	-	-	-	-	-
<b>Bare soil</b> ( $\omega_2$ )	-	100	96.80	-	-	-
<b>Vegetation</b> ( $\omega_3$ )	-	100	100	-	-	-
<b>Road</b> ( $\omega_4$ )	-	-	-	97.04	-	-
<b>Shadow</b> ( $\omega_5$ )	95.92	100	75.76	-	94.08	-
<b>Building</b> ( $\omega_6$ )	-	-	-	-	-	94.36

TABLE VI. AVERAGE (ON 10 TRIALS) ACCURACIES (%) OF EACH LAND-COVER TRANSITION OBTAINED BY AL-PCC (TRENTO DATA SET).

<b>July 2006</b> <b>October 2005</b>	<b>Water</b> ( $v_1$ )	<b>Bare soil</b> ( $v_2$ )	<b>Vegetation</b> ( $v_3$ )	<b>Road</b> ( $v_4$ )	<b>Shadow</b> ( $v_5$ )	<b>Building</b> ( $v_6$ )
<b>Water</b> ( $\omega_1$ )	99.59	-	-	-	-	-
<b>Bare soil</b> ( $\omega_2$ )	-	99.56	96.31	-	-	-
<b>Vegetation</b> ( $\omega_3$ )	-	99.95	100	-	-	-
<b>Road</b> ( $\omega_4$ )	-	-	-	87.35	-	-
<b>Shadow</b> ( $\omega_5$ )	96.66	100	84.03	-	93.01	-
<b>Building</b> ( $\omega_6$ )	-	-	-	-	-	80.46

TABLE VII. AVERAGE (ON 10 TRIALS) OVERALL ACCURACIES (%) PROVIDED BY JEAL AND AL-PCC UNDER DIFFERENT MISREGISTRATION CONDITIONS (TRENTO DATA SET).

<b>Amount of Misregistration</b> ( $n,m$ )	<b>Overall Accuracy (%)</b>	
	<b>JEAL</b>	<b>AL-PCC</b>
(0,0)	97.42	94.45
(+1,0)	96.02	92.68
(-1,0)	96.20	92.97
(0,+1)	95.92	93.30
(0,-1)	96.78	93.71
(+1,+1)	94.78	92.26

### B. Sardinia Data Set

Fig. 5 shows the behavior of the average (on 10 trials) overall accuracies obtained by the proposed AL algorithms in the detection of land-cover transitions versus the number of training samples on the Sardinia data set. By analyzing the figure, one can observe that the JEAL algorithm, again, provides the highest accuracies for most of the iterations and for both values of  $h$ . Moreover, it

reaches convergence in a smaller number of iterations than the other methods. In this case, the  $\text{JEAL}_{\text{Ind}}$  algorithm provides slightly higher accuracies than the MEAL one for both values of  $h$ . The JEAL algorithm achieves an accuracy of 95.53% with only 170 samples, whereas the  $\text{JEAL}_{\text{Ind}}$  and the MEAL techniques could not obtain such an accuracy. The accuracy yielded using the full pool as training set (2249 samples) is 93.22%, whereas the highest accuracies obtained by JEAL,  $\text{JEAL}_{\text{Ind}}$  and MEAL are 95.53% (with 170 samples), 94.81% (with 260 samples), and 94.84% (with 280 samples), respectively. The reason of achieving higher accuracies with JEAL,  $\text{JEAL}_{\text{Ind}}$  and MEAL compared to the case of using the whole pool as training set is related to the presence of noisy samples (or outliers) in the pool. These samples do not properly model the distribution of test pixels. It is worth nothing that an outlier is expected to be assigned to a wrong class by the classifier with high confidence (*i.e.*, with low uncertainty); accordingly, it is not selected as an uncertain sample by the proposed AL method. From all the results, we can observe that, when the compound classification is considered, the uncertainty criterion based on joint entropy that considers temporal dependence between multitemporal images (*i.e.*, JEAL) can improve the classification accuracy with respect to the standard marginal entropy (*i.e.*, MEAL) as well as the sum of entropies (*i.e.*,  $\text{JEAL}_{\text{Ind}}$ ). This demonstrates the importance of information conveyed by temporal dependence between multitemporal images for optimizing the definition of the training sets.

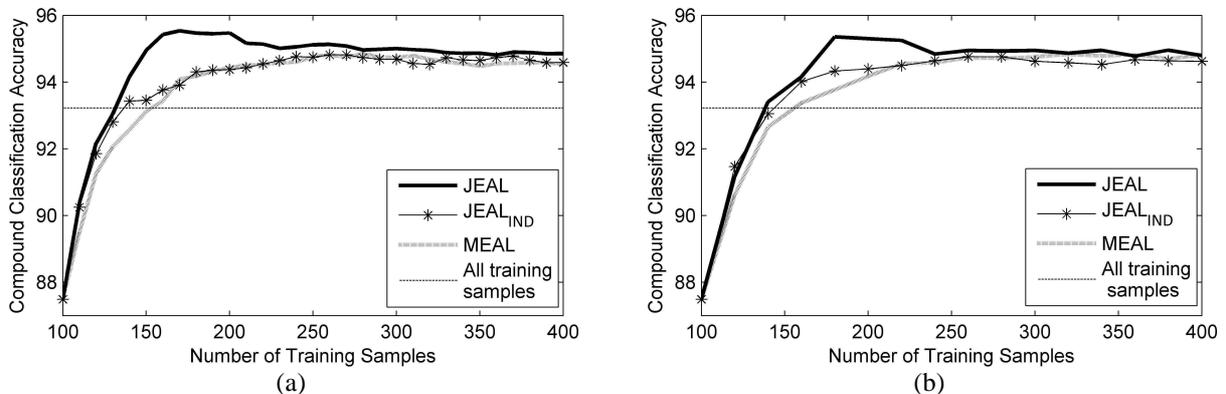


Fig. 5. Average (on 10 trials) overall accuracy in the detection of land-cover transitions obtained by using the JEAL,  $\text{JEAL}_{\text{Ind}}$  and MEAL algorithms when (a)  $h=10$  and (b)  $h=20$  and the full pool as training set (Sardinia data set).

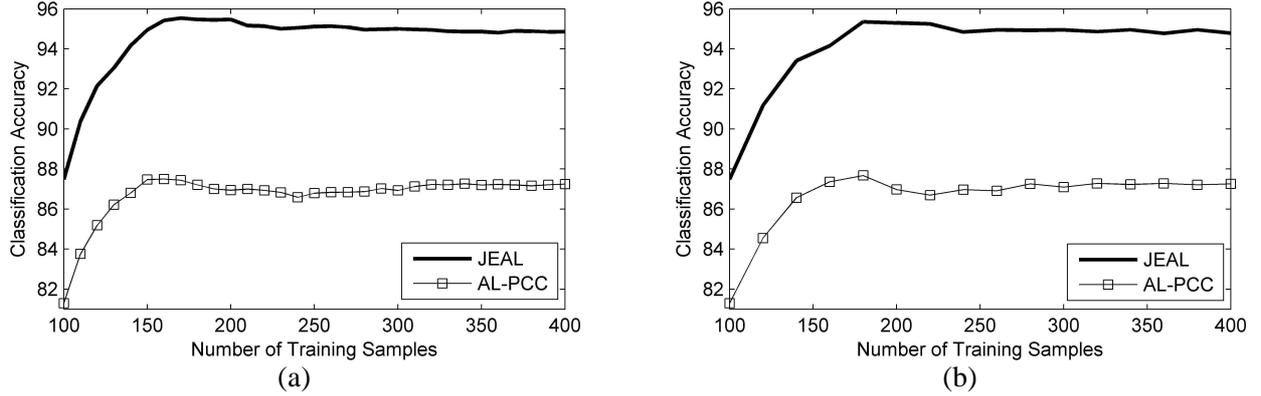


Fig. 6. Average (on 10 trials) overall accuracy in the detection of land-cover transitions obtained by the JEAL with compound classification (JEAL) and AL with post-classification comparison (AL-PCC) when (a)  $h=10$  and (b)  $h=20$  (Sardinia data set).

Fig. 6 shows the comparison of the accuracies in the detection of land-cover transitions obtained by JEAL and AL-PCC for both values of  $h$ . One can observe that the JEAL algorithm results in the highest accuracy at each iteration when compared to the AL-PCC method. These results confirm the effectiveness of both the compound classification and the proposed AL method based on the JEAL criterion. The average (on 10 trials) accuracies on the detection of each land-cover transition obtained by JEAL and AL-PCC are shown in TABLE VIII and Table IX, respectively, at the 7th iteration (*i.e.*, the training data includes 170 pairs of samples). From the tables, one can see that the accuracies obtained by AL-PCC is significantly improved by exploiting JEAL.

Table X shows the accuracies provided by both JEAL and AL-PCC at the 7th iteration (*i.e.*, the training data set includes 170 pairs of samples) by introducing misregistration between the images. The results are provided for different amounts of misalignment  $(n,m)$ . By analyzing the table, one can see that also on this data set the proposed method is always more accurate than the AL-PCC technique. As an example, in the case of  $(n,m)=(-1,0)$ , the accuracy of JEAL is 95.57%, whereas that of AL-PCC is 88.17%. It is worth noting that on this data set the accuracies obtained with misaligned images are not always lower than those yielded with co-registered images. This is mainly due to the fact that the available ground reference samples are mainly related to

homogeneous areas (*i.e.*, they are not close to the class borders), and thus the misregistration does not imply misclassification.

TABLE VIII. AVERAGE (ON 10 TRIALS) ACCURACIES (%) ON EACH LAND-COVER TRANSITION OBTAINED BY JEAL (SARDINIA DATA SET).

September 1995 \ July 1996	Pasture ( $v_1$ )	Forest ( $v_2$ )	Urban Area ( $v_3$ )	Water Body ( $v_4$ )	Wineyard ( $v_5$ )
	Pasture ( $\omega_1$ )	94.97	-	-	-
Forest ( $\omega_2$ )	-	94.39	-	-	-
Urban Area ( $\omega_3$ )	-	-	97.12	-	-
Water Body ( $\omega_4$ )	-	-	-	100	-
Wineyard ( $\omega_5$ )	-	-	-	-	64.95

TABLE IX. AVERAGE (ON 10 TRIALS) ACCURACIES (%) ON EACH LAND-COVER TRANSITION OBTAINED BY AL-PCC (SARDINIA DATA SET).

September 1995 \ July 1996	Pasture ( $v_1$ )	Forest ( $v_2$ )	Urban Area ( $v_3$ )	Water Body ( $v_4$ )	Wineyard ( $v_5$ )
	Pasture ( $\omega_1$ )	82.05	-	-	-
Forest ( $\omega_2$ )	-	92.26	-	-	-
Urban Area ( $\omega_3$ )	-	-	86.55	-	-
Water Body ( $\omega_4$ )	-	-	-	100	-
Wineyard ( $\omega_5$ )	-	-	-	-	47.26

TABLE X. AVERAGE (ON 10 TRIALS) OVERALL ACCURACIES (%) PROVIDED BY JEAL AND AL-PCC UNDER DIFFERENT MISREGISTRATION CONDITIONS (SARDINIA DATA SET).

Amount of Misregistration ( $n,m$ )	Overall Accuracy (%)	
	JEAL	AL-PCC
(0,0)	95.53	87.43
(+1,0)	95.16	87.47
(-1,0)	95.57	88.17
(0,+1)	95.98	86.96
(0,-1)	94.99	86.89
(+1,+1)	95.40	87.60

## VII. DISCUSSION AND CONCLUSION

In this paper, active learning for the detection of land-cover transitions in multitemporal remote sensing images has been addressed. This has been done by introducing active learning in the

framework of the compound classification and presenting a novel uncertainty criterion defined in the context of Bayes rule for compound classification. The proposed uncertainty criterion is based on the joint entropy associated with the compound classification decisions and is implemented under different simplifying assumptions on the temporal dependence between images. The first assumption considers the class-conditional independence in the time domain. This assumption is necessary because, as explained in the paper, in practice it is very difficult to define a training set suitable for a reliable estimation of the joint conditional posterior probabilities of all possible combinations of classes. Under this assumption, the uncertainty of a pair of samples is evaluated taking into account the temporal dependence modeled by the joint prior probabilities of classes (JEAL algorithm). The second stronger assumption introduces temporal independency among prior probabilities of classes together with the hypothesis of class-conditional independence in the time domain. This results in the calculation of the joint entropy of corresponding pairs of pixels as the sum of marginal entropies, *i.e.*, the mutual information between decisions is assumed to be zero (JEAL<sub>Ind</sub> algorithm). The third assumption also neglects the information associated with the spatial correspondence of pixels (MEAL algorithm). All the derived active learning algorithms have been theoretically and experimentally compared with each other. Experimental results show that using the temporal dependence in the definition of active learning for compound classification problems (*i.e.*, exploiting the definition of joint entropy obtained under the first assumption) results in higher accuracies in the detection of the land-cover transitions than the other algorithms when the same number of labeled samples are considered. From another perspective, the JEAL algorithm can achieve the same accuracy achieved by other techniques with a sharply smaller number of labeled samples. This is a very significant advantage, given the complexity and the cost of the collection of reference samples, especially in a multitemporal context. Summarizing, we can state that:

- 1) JEAL exploits in the query function the temporal dependence between images modeled by

both the prior joint probabilities of classes and the corresponding spatial position of pixels.

- 2) JEAL<sub>Ind</sub> exploits in the query function as only source of temporal information the corresponding spatial position of pixels (which can be misleading when the prior joint probabilities of classes are modeled as the product of the marginal prior probabilities of classes on each single image).
- 3) MEAL assumes complete independence between images in the query function.

In the experimental analysis we also compared the proposed active learning algorithms defined in the context of compound classification with a marginal entropy based active learning technique based on the post-classification comparison. By this comparison, we observed that the accuracies in the detection of land-cover transitions obtained by active learning with post-classification comparison are significantly increased by the proposed technique thanks to the information extracted from temporal dependence between images in both the active learning and the classification tasks. In addition, we also analyzed the effects of registration errors on the performance of both techniques. By this analysis, we observed that the residual misregistration between images may decrease the classification performance of both techniques. Nonetheless, the proposed algorithm still outperforms the AL-PCC also in presence of misregistration.

As a final remark, we would like to point out that the use of efficient techniques for the exploitation of supervised change-detection methods in real applications is becoming more and more important. This is due to the increased complexity of the first generation of the satellite VHR and hyperspectral images, which decreases the effectiveness of unsupervised change-detection methods. In this context, the proposed approach is very promising as it allows to optimize the definition of a multitemporal training set to be used in change detection, decreasing significantly the cost and effort required for multitemporal reference data collection.

As a future development of this work, we plan to extend the proposed active learning algorithms by including a diversity criterion defined in the context of compound classification.

This criterion should consider the concept of diversity in the selection of unlabeled pairs of samples, *i.e.*, it should select uncertain samples that are also diverse to each other to reduce the possible redundancy in the defined training set. In addition, we are also investigating the use of the proposed active learning algorithms when constraints on some available labels (*e.g.*, on the first image) are given.

## REFERENCES

- [1] K. Green, D. Kempka, and L. Lackey, "Using remote sensing to detect and monitor land-cover and land-use change," *Photogramm. Eng. Remote Sens.*, vol. 60, no. 3, pp. 331–337, 1994.
- [2] L. Bruzzone and D. F. Prieto, "Automatic analysis of the difference image for unsupervised change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 38, no. 3, pp. 1171–1182, May 2000.
- [3] F. Bovolo, "A multilevel parcel-based approach to change detection in very high resolution multitemporal images," *IEEE Geosci. Rem. Sens. Lett.*, vol. 6, no. 1, pp. 33–37, Jan. 2009.
- [4] F. Bovolo and L. Bruzzone, "A theoretical framework for unsupervised change detection based on change vector analysis in polar domain," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 1, pp. 218–236, Jan. 2007.
- [5] A. Singh, "Digital change detection techniques using remotely-sensed data," *Int. J. Remote Sensing*, vol. 10, no. 6, pp. 989–1003, 1989.
- [6] L. Bruzzone, and S.B. Serpico, "An iterative technique for the detection of land-cover transitions in multitemporal remote-sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 35, no. 4, pp. 858-867, July 1997.
- [7] L. Bruzzone, D. Fernandez Prieto, and S.B. Serpico, "A neural-statistical approach to multitemporal and multisource remote-sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, no.3, pp. 1350-1359, May 1999.
- [8] L. Bruzzone, R. Cossu, and G. Vernazza, "A neural-statistical approach to multitemporal and multisource remote-sensing image classification," *Pattern Recognition Letters*, vol. 25, pp. 1491–1500, 2004.
- [9] L. Bruzzone, and F. Bovolo, "A conceptual framework for change detection in very high resolution remote sensing images," *IEEE International Geoscience and Remote Sensing Symposium, Hawaii, USA, pp. 2555-2558, 2010*
- [10] D. Lewis and W. Gale, "A sequential algorithm for training text classifiers," *In Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3–12. ACM/Springer, 1994.
- [11] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *In Proceedings of the International Conference on Machine Learning*, Stanford, USA, pp. 999–1006, 2000.
- [12] C. Campbell, N. Cristianini, and A. Smola, "Query learning with large margin classifiers", *Proc. 17<sup>th</sup> Int'l Conf. Machine Learning*, Stanford, USA, pp. 111-118, 2000.
- [13] Z. Xu, K. Yu, V. Tresp, X. Xu, and J. Wang, "Representative sampling for text classification using support vector machines," *25th European Conf. on Information Retrieval Research*,

Pisa, Italy, pp. 393-407, 2003.

- [14] K. Brinker, "Incorporating diversity in active learning with support vector machines," *Proceedings of the International Conference on Machine Learning*, Washington DC, pp. 59-66, 2003.
- [15] C. Zhang and T. Chen, "An active learning framework for content-based information retrieval," *IEEE Transactions on Multimedia*, vol. 4, no. 2, pp. 260-268, 2002.
- [16] J. Zhu, and E. Hovy, "Active learning for word sense disambiguation with methods for addressing the class imbalance problem," *In Proc. Joint Conf. Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Prague, 783-790, 2007.
- [17] A. McCallum and K. Nigam, "Employing EM in pool-based active learning for text classification," *In Proceedings of the International Conference on Machine Learning*, Madison, Wisconsin USA, pp 359-367, 1998.
- [18] I. Dagan and S. Engelson, "Committee-based sampling for training probabilistic classifiers," *In Proceedings of the International Conference on Machine Learning*, pp. 150-157, 1995.
- [19] P. Mitra, B. U. Shankar, and S. K. Pal, "Segmentation of multispectral remote sensing images using active support vector machines," *Pattern Recognit. Lett.*, vol. 25, no. 9, pp. 1067-1074, Jul. 2004.
- [20] Q. Liu, X. Liao, and L. Carin, "Detection of unexploded ordnance via efficient semisupervised and active learning", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 9, pp. 2558-2567, Sept. 2008.
- [21] S. Rajan, J. Ghosh, and M. M. Crawford, "An active learning approach to hyperspectral data classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 4, pp. 1231-1242, Apr. 2008.
- [22] D. Tuia, F. Ratle, F. Pacifici, M. Kanevski, and W. J. Emery, "Active learning methods for remote sensing image classification," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 47, no. 7, pp. 2218 -2232, Jul. 2009.
- [23] A. Liu, G. Jun and J. Ghosh, "A self-training approach to cost sensitive uncertainty sampling," *Machine Learning Journal*, vol. 76, no. 2-3, pp. 257-270, Sept. 2009.
- [24] A. Liu, G. Jun, and J. Ghosh, "Active learning of hyperspectral data with spatially dependent label acquisition costs," *IEEE International Geoscience and Remote Sensing Symposium*, Cape Town, South Africa, pp. V-256 - V-259, 2009.
- [25] B. Demir, C. Persello, and L. Bruzzone, "Batch mode active learning methods for the interactive classification of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no.3, pp. 1014-1031, March 2011.
- [26] S. Patra and L. Bruzzone, "A fast cluster-based active learning technique for classification of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, 2011, in press.
- [27] L. Bruzzone and C. Persello, "Recent trends in classification of remote sensing data: active and semisupervised machine learning paradigms," *IEEE International Geoscience and Remote Sensing Symposium*, Hawaii, USA, pp.3720-3723, 2010.
- [28] Bruzzone, M. Chi, and M. Marconcini, "A Novel Transductive SVM for the Semisupervised Classification of Remote-Sensing Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 11, pp. 3363-3373, Nov. 2006.
- [29] G. Camps-Valls, T.V. Bandos, Marsheva, and D. Zhou, "Semi-supervised graph-based hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 10, pp. 3044-3054, Oct. 2007.

- [30] L. Bruzzone and C. Persello, "A novel context-sensitive semisupervised SVM classifier robust to mislabeled training samples," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 7, pp. 2142-2154, July 2009.
- [31] A. Papoulis and S. U. Pillai, *Probability, Random Variables, and Stochastic Processes*, McGraw Hill, New York, 2002.
- [32] J.A. Richards and X. Jia, *Remote Sensing Digital Image Analysis: an Introduction*, Springer-Verlag, New York, 1999.