Title: A Novel Active Learning Method in Relevance Feedback for Content Based Remote Sensing Image Retrieval

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

Authors: B. Demir, L. Bruzzone,

Volume: 53, Issue: 5, May 2015

Page(s): 2323 – 2334

DOI: 10.1109/TGRS.2014.2358804

A Novel Active Learning Method in Relevance Feedback for Content Based Remote Sensing Image Retrieval

Begüm Demir, *Member IEEE*, 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, lorenzo.bruzzone@ing.unitn.it.

***Abstract*—**Conventional relevance feedback (RF) schemes improve the performance of content-based image retrieval (CBIR) requiring the user to annotate a large number of images. To reduce the labeling effort of the user, this paper presents a novel active learning (AL) method to drive RF for retrieving remote sensing images from large archives in the framework of Support Vector Machine classifier. The proposed AL method is specifically designed for CBIR and defines an effective and as small as possible set of relevant and irrelevant images with regard to a general query image by jointly evaluating three criteria: i) uncertainty, ii) diversity and iii) density of images in the archive. The uncertainty and diversity criteria aim at selecting the most informative images in the archive, whereas the density criterion goal is to choose the images that are representative of the underlying distribution of data in the archive. The proposed AL method assesses jointly the three criteria based on two successive steps. In the first step the most uncertain (i.e., ambiguous) images are selected from the archive on the basis of margin sampling strategy. In the second step the images that are both diverse (i.e., distant) to each other and associated to high density regions of the image feature space in the archive are chosen from the most uncertain images. This step is achieved by a novel clustering based strategy. The proposed AL method for driving the RF contributes to mitigate problems of unbalanced and biased set of relevant and

irrelevant images. Experimental results show the effectiveness of the proposed AL method.

Index Terms **–** active learning, content based image retrieval, relevance feedback, remote sensing.

I. INTRODUCTION

With the development of satellite technology large-volume remote sensing (RS) images (i.e., millions of single-date as well as time-series of Earth observation scenes) becomes available. Accordingly, one of the most challenging and emerging applications in RS is the efficient and precise retrieval of RS images from such archives according to the users' needs. Conventional remote sensing image retrieval systems often rely on keywords/tags in terms of sensor type, geographical location and data acquisition time of images stored in the archives. The performance of tag matching based retrieval approaches highly depends on the availability and the quality of manual tags. However, in practice keywords/tags are expensive to obtain and often ambiguous. Due to these drawbacks, recent studies have shown that the content of the RS data is more relevant than manual tags. Accordingly, content-based image retrieval (CBIR) has attracted increasing attentions in the RS community particularly for its potential practical applications to RS image management. This will become particularly important in the next years when the number of acquired images will dramatically increase. Any CBIR system essentially consists of (at least) two modules [1],[2]: i) a feature extraction module that derives a set of features for characterizing and describing images, and ii) a retrieval module that searches and retrieves images similar to the query image. Querying image contents from large RS data archives depends on the capability and effectiveness of the feature extraction techniques in describing and representing the images. In the RS literature, several primitive (i.e., low-level) features have been presented for retrieval purposes, such as: intensity features [5]; color features [6],[7]; shape features[8]-[10], texture features [10]-

[16]; and local invariant features [17]. However, the low-level features from an image have very limited capability in representing and analyzing the high-level concept conveyed by RS images (i.e., the semantic content of RS images). This issue is known as semantic gap occurred between the low-level features and high-level semantic content, and leads to poor CBIR performance. Consequently, the semantic gap is the crucial challenge in CBIR applications.

In order to confine the semantic gap, relevance feedback (RF) schemes have been designed to iteratively improve the performance of CBIR by taking user's (i.e., an oracle who knows the correct labeling of all images) feedback into account [3], [4]. At each iteration, the user's feedback is used to provide relevant and irrelevant images to the query image that are positive and negative feedback samples, respectively. RF can be considered as a binary-classification problem: one class includes relevant images, and the other one consists of the irrelevant ones. Then, any supervised classification method can be used in the context of CBIR by training the classifier with the already annotated images of two-classes [3], [4]. Accordingly, during RF, the search strategy is refined iteration by iteration by improving the classification model with the recently annotated images. As mentioned above, user involvement is required at each RF iteration for annotating images. However, labeling images as relevant or irrelevant is time consuming and thus costly. Accordingly, despite the retrieval success of RF, the conventional RF schemes are not practical and efficient in real applications, especially when huge archives of remote sensing images are considered.

An effective approach to reduce the annotation effort in RF is active learning (AL) that aims at finding the most informative images in the archive that, when annotated and included in the set of relevant and irrelevant images (i.e., the training set) can significantly improve the retrieval performance [10]. Moreover, selecting the most informative images results in: i) a smaller number of RF iterations to optimize the CBIR, and ii) a reduced annotation time due to the optimization of the training set with a minimum number of highly informative images. In the RS community most

of the previous studies in AL have been developed in the context of classification problems for land-cover maps generation (see [18] for comprehensive review on the most relevant techniques). In particular the unlabeled samples that are highly uncertain and diverse to each other are usually selected as informative samples to be labeled and included in the training set for the classification of RS [18]. The uncertainty of a sample is related to the confidence of the supervised algorithm in correctly classifying it, whereas the diversity among samples is associated to their correlation in the feature space (i.e., samples that are as distant as possible to each other are the most diverse samples).

From the AL perspective the CBIR problem is more complex than the standard classification problem due to the fact that: i) in general the class of irrelevant images (which is dynamically driven on the basis of the specific query image given as input to the classifier) is much larger than the class of relevant images because the irrelevant class consists of the huge number of images that in a real archive are irrelevant to the query image, ii) the classifier is trained with a largely incomplete number of annotated images (training set) due to the absence of many irrelevant image categories (those that existing in the archive) within the training set and iii) in real large-scale RS archives the total number of images is usually very large. All the above-mentioned reasons result in strongly imbalanced and biased training sets. As a result, the boundary between two classes is initially unstable and inaccurate, and thus it does not allow a reliable modeling of the problem. Accordingly, AL methods defined for classification problems that only assess uncertainty and diversity of samples are not efficient for CBIR problems.

AL has been marginally considered in the framework of CBIR problems in the RS community. At the best of our knowledge only one AL method is presented [10], which is developed in the context of Support Vector Machine (SVM) classifier and inspired from AL methods used for classification problems [21]. In this method the uncertainty and diversity criteria have been applied in two consecutive steps. In the first step the most uncertain images are selected

from the archive. To this end, the unlabeled images closest to the current separating hyperplane (those that are the most uncertain) are initially selected by margin sampling (MS) [19],[20]. In the second step, the images that are diverse to each other among the uncertain ones are chosen on the basis of the distances estimated between them. An important shortcoming of the method presented in [10] is that it does not evaluate the representativeness of images in terms of their density in the archive. However, images that fall into the high density regions of the image feature (descriptor) space are crucial for CBIR problems particularly when a small number of initially annotated images is available. This is due to the fact that they are statistically very representative of the underlying image distribution in the archive. Therefore, the retrieval results on them affect much more the overall retrieval accuracy than the results obtained on images within low density regions.

To overcome the above-mentioned critical issues, in this paper we propose a CBIR approach that include a novel triple criteria AL (TCAL) method to drive RF in CBIR. For the selection of the most informative as well as representative unlabeled images of images to be annotated, the proposed TCAL method jointly evaluates three criteria: i) relevancy, ii) diversity and iii) density of images in the archive. In order to assess the above-mentioned three criteria, the proposed TCAL method exploits a two-step procedure defined in the framework of the SVM classifier. In the first step, the most uncertain (i.e., ambiguous) images are selected by the well-known MS strategy [10], whereas in the second step the diverse images among the most uncertain ones are selected from the highest density regions of image feature space. The latter step is achieved by a novel clustering based strategy that evaluates the density and diversity of unlabeled images in the image feature space to drive the selection of images to be annotated.

The novelties of the proposed AL method for RF in CBIR consist in: 1) the design and development of a strategy to jointly evaluate the three criteria (i.e., relevancy, diversity and density) for the selection of most informative and representative images in the context of CBIR problems; 2) the use of the prior term of the distributions based on the density of unlabeled images

in the image feature space to assess the representativeness of images and thus to identify images to annotate. Thanks to the joint use of three criteria and to the use of the density of images in the image feature space, the proposed TCAL method can effectively avoid various problems caused by insufficient number of annotated samples in RF, and thus is appropriate and effective for RS image retrieval. Moreover, we introduce the use of histogram intersection (HI) kernel in the RS community in the framework of the proposed CBIR approach as a similarity measure of image features in the kernel space. Note that in recent years the HI kernel has gained an increasing interest for image retrieval problems in the computer-vision communities [25], [26], [27], whereas its use in RS has not been explored yet. Experiments carried out on an archive of aerial images demonstrate the effectiveness of the proposed method.

The remaining part of this paper is organized as follows. Section II introduces the proposed AL method. Section III describes the considered data archive and the design of CBIR system, whereas Section IV illustrates the experimental results. Finally, Section V draws the conclusion of this work.

II. PROPOSED METHOD

A. Problem Formulation

Let us consider an archive Υ made up a very large number of R remote sensing images $\{\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_R\}$, where $\mathbf{X}_i$ is the i-th image defined as $\{x_i^1, x_i^2, ..., x_i^L\}$, $i = 1, ..., R$. x_i^l, $l = 1, ..., L$, is the l-th feature characterizing the content of the i-th image in Υ and L is the total number of features. Let $\mathbf{X}_q = \{x_q^1, x_q^2, ..., x_q^L\}$ be a query image that can be selected by the user from the archive Υ (i.e., $\mathbf{X}_q \in \Upsilon$) or outside the archive Υ (i.e., $\mathbf{X}_q \notin \Upsilon$). A general CBIR system with RF driven by AL consists of three modules: 1) primitive (low-level) feature extraction module that is applied to both query image and all images in the archive; 2) initial training set definition module that builds an initial training set T with a small number relevant and irrelevant images with respect

to query; and 3) RF driven by an AL module that enriches the training set T defined by the previous module, and returns the set δ of images from the archive Υ. Fig. 1 shows the general block scheme of the CBIR with RF driven by AL. In this paper, we mainly focus on the RF driven by AL module (see Fig. 2) which is a crucial part for the success of the CBIR system. Then we briefly address the feature extraction module and the important choices adopted for assessing the similarities of image features in the proposed system.

AL iteratively expands the size of an initial labeled training set T selecting the most informative images from the archive Υ for their annotation. At each RF iteration, the most informative unlabeled images for a given classifier are: i) selected based on an AL function, ii) annotated by a supervisor (i.e., an oracle) and iii) added to the current training set T. Finally, the supervised classifier is retrained with the images moved from Υ to T. It is worth nothing that the initial training set T requires few annotated images for the first training of the classifier and then is enriched iteratively by including the most informative images selected from Υ. At each iteration, after the classifier is trained, the retrieval of the images under investigation is carried out. These processes are repeated until the user is satisfied with retrieval results. The general flowchart of the AL based RF approach is given in Fig. 2. The selection of the most informative samples from Υ to be included in the training set T on the basis of AL offers two main advantages: i) the annotation cost is reduced due to the avoidance of redundant images; and ii) an accurate retrieval 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 (images to be annotated are selected from the classifier as the most informative for its classification rule). Of course the success of the RF strongly depends on the capability of the specific AL method considered to select the most informative and representative images to be annotated in order to limit as much as possible the effort of the user for reaching the final relevant result.

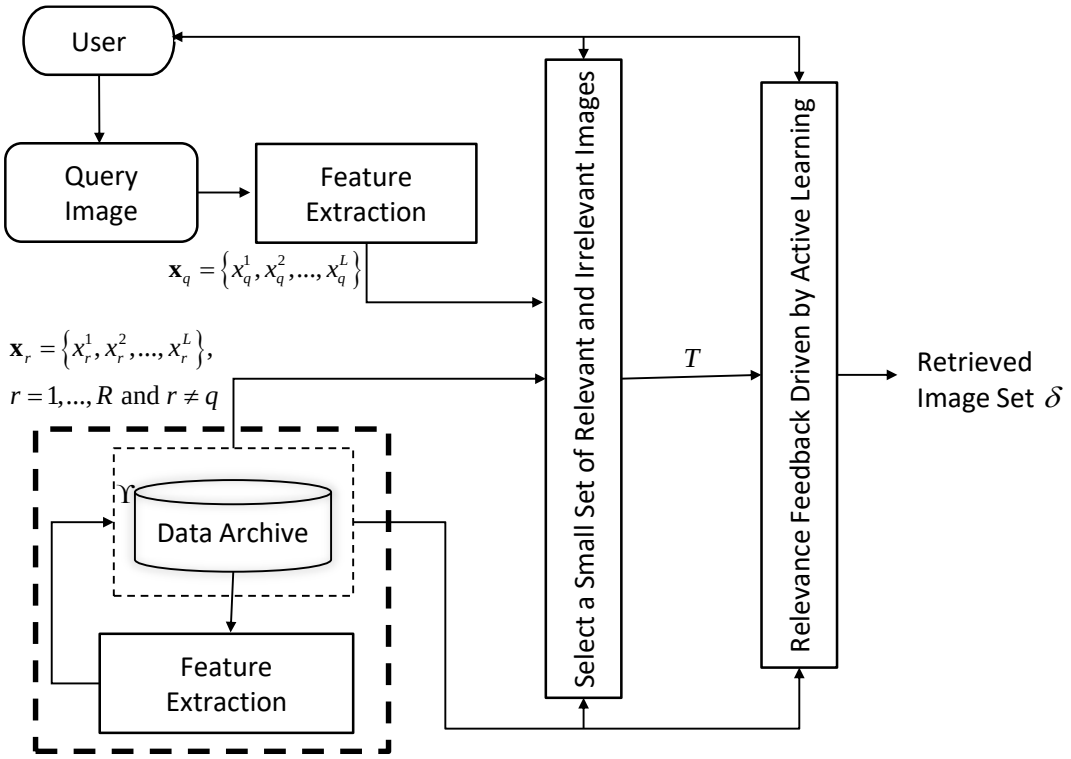


Fig. 1. The general architecture of a CBIR system with RF driven by AL

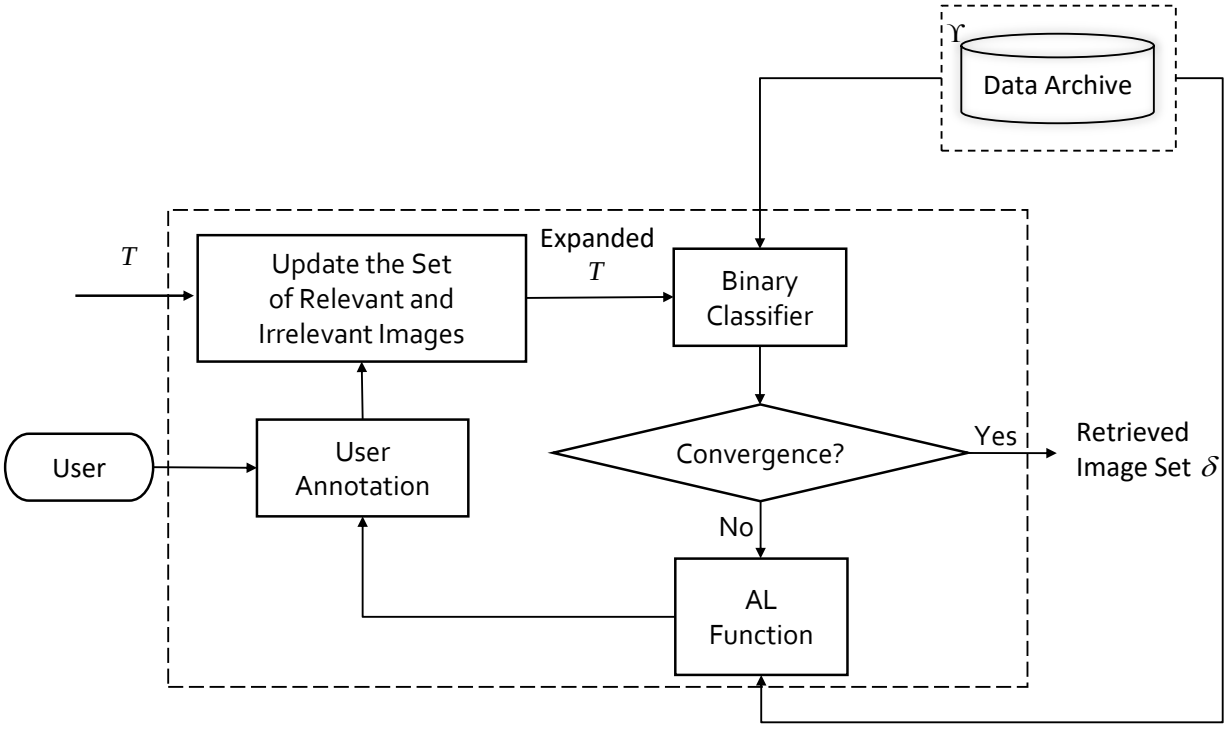


Fig. 2. General flowchart of RF driven by AL.

B. Proposed Triple Criteria Active Learning Method

We propose a novel triple criteria AL (TCAL) method to expand the initial training set during RF rounds in CBIR applications. The aims of the proposed AL method are as follows: i) to achieve a training set of annotated relevant and irrelevant images with respect to the query image as small as possible within a low number of RF iterations; and ii) to retrieve the images similar to the query image with high accuracy. The proposed TCAL method is defined in the context of binary Support Vector Machine (SVM) classification and selects a batch $S = \{\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_h\}$ of h images at each RF iteration that are i) uncertain (i.e., ambiguous), ii) as more diverse as possible to each other, and iii) located in the highest density regions of image feature space. The uncertainty of images is assessed according to the MS strategy, whereas diversity and density of image are evaluated by a novel clustering based strategy. At each iteration, the proposed AL method jointly evaluates the above-mentioned three criteria by a strategy that is based on two consecutive steps to select the batch S of images. In the first step, the $m > h$ most uncertain images are selected according to the standard MS technique from Υ. In the second step the most diverse h images among these m uncertain (i.e., ambiguous) images are chosen from highest density regions of the feature space

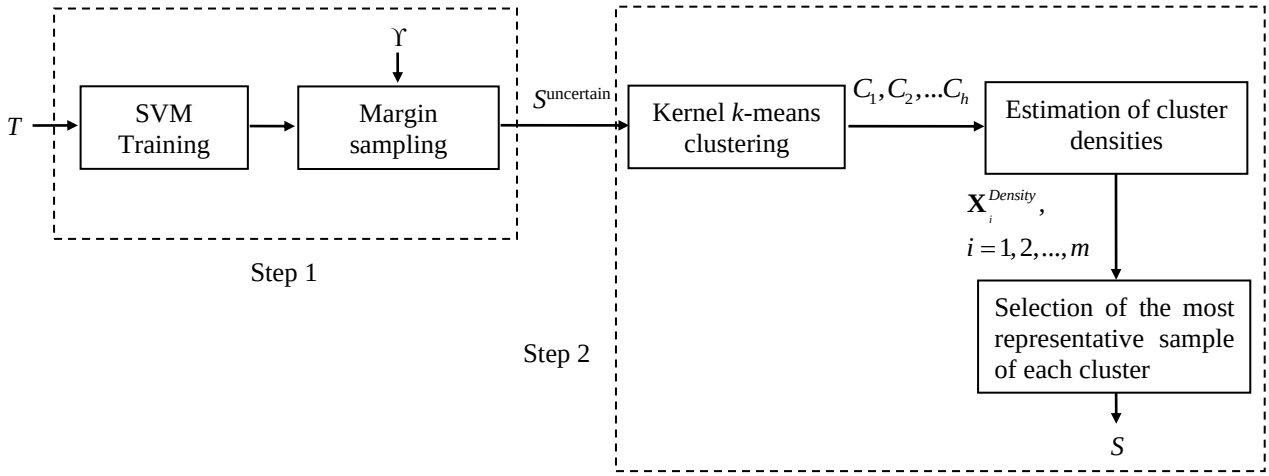


Fig. 3. Block diagram of the proposed AL method

($m > h > 1$). Fig. 3 shows the general block scheme of the proposed AL method. The first step is devoted to select unannotated images that have maximum uncertainty on their correct target classes according to the binary SVM classification properties. The basic idea behind this concept is that images, which have the lowest probability to be accurately classified by the considered classifier, are the most beneficial to be included in the training set for separating the two categories of relevant and irrelevant images in an optimal way. For SVM classification the images closest to the separating hyperplane (which is the discriminant function) have low confidence to be correctly classified. One of the most popular AL method in the context of SVM classification is MS, which selects the unlabeled samples closest to the separating hyperplane, as they are the samples considered with the lowest confidence (i.e., those that have the maximal uncertainty on the true information class). Accordingly, we considered this approach to select the most uncertain images in the first step due to its simplicity and effectiveness and possible fast implementation. To this end, initially a binary SVM is trained using the existing set of relevant and irrelevant images. Then, the functional distances of the unannotated images to the current SVM hyperplane are estimated. The set of *m* images $S^{\text{uncertain}} = \{\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_m\}$ (where $\mathbf{X}_i = \{x_i^1, x_i^2, ..., x_i^L\}$ is the *i*-th uncertain image described by *L* primitive features) closest to the corresponding separating hyperplane are selected. It is worth noting that the selection of the value of *m* is important for the effectiveness of the proposed AL method. Fixing *m* as a very high value may result in the selection of images with a low degree of certainty, whereas defining very small *m* values may cause neglecting highly uncertain images. Thus *m* should be defined carefully and according to previous studies carried out in the AL literature [18] we define it as $m = 4h$.

The second step is devoted to select *h* images from the set $S^{\text{uncertain}}$ of the most uncertain images that are diverse to each other, taking into account sample density in the archive image feature space. Selecting the uncertain images from high density regions of the image feature space is crucial in the proposed TCAL method. This is due to the fact that, under the reasonable

assumption that images in the same region of the image feature space have similar target categories, the selection of images to be annotated from high density regions is an effective strategy for minimizing the overall retrieval error. This choice aims to reduce errors in regions of the image feature space where we have many uncertain unlabeled images that can strongly affect the overall retrieval accuracy.

To analyze the density and diversity of uncertain images, in this step we propose a novel clustering-based approach. The use of clustering is due to the fact that it is an effective way to evaluate the diversity and density of images, since 1) unlabeled uncertain images from different clusters are implicitly sparse in the feature space, thus they can be considered as diverse images, and 2) density of each cluster can give information on the density of the images in the associated region of the image feature space. Although any clustering technique can be exploited, here we select the kernel *k*-means clustering technique, since 1) it operates in the kernel space where the SVM separating hyperplane is defined, and 2) it is proven to be more effective to identify non-linearly separable clusters in the case of non-linearly separable data than conventional clustering (e.g., k-means) techniques [22], [23]. Accordingly, the kernel *k*-means clustering is initially applied to the set $S^{\text{uncertain}} = \{\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_m\}$ of uncertain images, and these images are divided into $k=h$ clusters ($C_1, C_2, ...C_h$) in the kernel space. Then the density $\mathbf{X}_i^{Density}$ associated with each image $\mathbf{X}_i$, $i = 1,2,...,m$ in the kernel space is estimated by computing the average distance between $\mathbf{X}_i$ and all the other images located in the same cluster. Let us assume that the image $\mathbf{X}_i$ falls within the cluster $C_v, v = 1,2,...,k$. Then, the density $\mathbf{X}_i^{Density}$ of $\mathbf{X}_i$ is estimated as:

$$\mathbf{X}_i^{Density} = \frac{1}{|C_v|} \sum_{\forall \mathbf{X}_j \in C_v} D^2(\phi(\mathbf{X}_i), \phi(\mathbf{X}_j)) = \frac{1}{|C_v|} \sum_{\forall \mathbf{X}_j \in C_v} \left\| \phi(\mathbf{X}_i) - \phi(\mathbf{X}_j) \right\|^2 \quad (1)$$

where $|C_v|$ is the total number of samples in C_v, $D^2(.,.)$ is the Euclidean distance between two images in the image feature space, and $\phi(\cdot)$ is a nonlinear mapping function from the original

image feature space to a higher dimensional space. The distance in the higher dimensional image feature space can be estimated by using only the kernel function $K(\cdot,\cdot)$ (see the next sub-section for the details on the considered kernel function in this paper) without considering the direct knowledge of the mapping function $\phi(\cdot)$, i.e., :

$$\left\|\phi\left(\mathbf{X}_i\right)-\phi(\mathbf{X}_j)\right\|^2 = K\left(\mathbf{X}_i,\mathbf{X}_i\right)-2K\left(\mathbf{X}_i,\mathbf{X}_j\right)+K\left(\mathbf{X}_j,\mathbf{X}_j\right) \quad (2)$$

Then, the sample that is the most representative of the underlying image distribution within the considered cluster is selected. Accordingly, after estimating the density of images in C_v, the image $\mathbf{X}_v$ that is associated with the portion of the image feature space with the highest density (i.e., the smallest average distance) in the cluster C_v is selected as the most representative image, i.e.,

$$\mathbf{X}_v = \underset{\forall \mathbf{X}_j \in C_v}{\arg\max}\left\{\mathbf{X}_j^{Density}\right\} \quad (3)$$

A set $S=\left\{\mathbf{X}_1,\mathbf{X}_2,...,\mathbf{X}_h\right\}, S\subset S^{uncertain}$ of *h* samples are selected from the *h* clusters (one for each cluster). Due to selection of one sample from each cluster, the diversity of images extracted at each iteration is achieved. At the end of this task a new training set with annotated images is obtained as $T=T\bigcup S$. It is worth noting that the joint use of the three criteria results in the selection of informative (as a result of uncertainty and diversity criteria) and representative (as a result of density criterion) images to annotate. The steps of AL are iterated until either the desired number of images is annotated or the retrieval accuracy satisfies user's requirements.

C. RS Image Feature Extraction and Classification in the Context of CBIR

We model RS images by exploiting a bag-of-visual-words (BOVW) representation of the local invariant features extracted by the scale invariant feature transform (SIFT). The SIFT is a translation, rotation and scale invariant image feature extraction technique and has recently been found very effective and robust in the context of RS image retrieval [17]. The SIFT results in

various local interest points within an image and their descriptors (i.e., SIFT descriptors) that characterize portions of images around the interest points. In order to summarize the SIFT descriptors by the bag-of-visual-words representation (that is generally considered for the local image descriptors), we apply kernel *k*-means clustering to a subset of randomly selected SIFT descriptors. This process results in a code-book. Then, the descriptors extracted from each image are quantized by assigning the label of the closest cluster [17]. Accordingly, the final representation of an image is the histogram (i.e., frequency) of the codebook entries (known as code-words) in the image [17]. Note that the histogram-based image representation is very popular for the bag-of-visual-words approaches that result to be the state-of-the-art in many image retrieval problems outside remote sensing.

In order to assess the similarities of the BOVW representations (histogram-based features) of the images in the kernel space we introduce in RS the use of histogram intersection (HI) kernel. Note that the similarity is used in both the SVM classification and the proposed AL method. To measure the similarities between the images $\mathbf{X}_i = \left\{x_i^1, x_i^2, ..., x_i^L\right\}$ and $\mathbf{X}_j = \left\{x_j^1, x_j^2, ..., x_j^L\right\}$, the HI kernel is defined as:

$$K\left(\mathbf{X}_i, \mathbf{X}_j\right) = \sum_{l=1}^{L} \min\left(x_i^l, x_j^l\right) \tag{4}$$

where each component $x_i^l \in \mathbf{X}_i, l = 1, 2, ..., L$ and $x_j^l \in \mathbf{X}_j, l = 1, 2, ..., L$ denotes a histogram feature. Note that the HI kernel is a positive definite parameter-free kernel for non-negative features (see [25]) and it has been recently found very effective in various computer-vision tasks (where histograms are popular representations of images) such as content-based image retrieval [25],[26], [27].

III. Data Set Description and Set up of the System

In order to assess the effectiveness of the proposed AL method we carried out several experiments on an archive that consists of images characterizing 21 categories (i.e., classes) selected from aerial orthoimagery [17]. Each category includes 100 images that were downloaded from the USGS National Map of the following US regions: Birmingham, Boston, Buffalo, Columbus, Dallas, Harrisburg, Houston, Jacksonville, Las Vegas, Los Angeles, Miami, Napa, New York, Reno, San Diego, Santa Barbara, Seattle, Tampa, Tucson, and Ventura. Each image in the archive is a section of 256×256 pixels with a spatial resolution of 30 cm, and belong to one of the following 21 classes: agriculture, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium density residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, and tennis courts. Fig. 4 shows example of two images for each category. For the further detailed information on the archive, we refer to the reader to [17]. Note that this archive is a benchmark and thus we know that we have 21 categories and all the images are already annotated. However, we use this archive simulating a real scenario where we do not know anything in the initial phase on the archive and we just use an image query for searching similar images also without identifying a specific category. This is an important observation that should be understood in order to avoid confusion between the addressed CBIR problem and standard supervised classification for the production of thematic maps. Another important point to emphasize is that for obvious reasons the benchmark is composed of a moderate number of images since for performance assessments we need annotations. In real applications the search is expected to be applied to much larger archives.

In the experiments, in order to obtain the BOVW representations of images (which summarizes the SIFT descriptors), kernel k-means clustering was applied to 100000 randomly selected SIFT descriptors by selecting k=150. Then, the SIFT descriptors are quantized by assigning the label of the closest cluster. The images downloaded from the National Map are in the

red-green-blue (RGB) color space. In order to use SIFT a coherent way with [17], each image is converted to grayscale. In the experiments, L2 normalized SIFT histogram features have been used, i.e., the components are normalized so that the feature vectors have length one.

Fig. 4. Example of two images for each category in the considered archive.

In the experiments, the value of the regularization parameter *C* of SVM was obtained by a five-fold cross validation implicitly done on the annotated images at each AL iteration. We carried out many different experiments. In the paper, for space constraints we report and discuss in detail the cases in which the initial query image is selected from the categories of: i) forest; and ii) agriculture. Then, in order to give a general overview of the performance of the proposed method we provide a summary of the results obtained by extracting the query image from all the 21 different categories present in the archive. All experimental outcomes are referred to the average results obtained in 30 trials according to thirty randomly selected initial query images from each category. To define the initial training set two relevant and three irrelevant images are randomly selected by the user. This choice results in a poor and imbalanced initial set of annotated images (i.e., initial training set) due to i) the small number of images used, and ii) the absence of images of many categories.

We compared the proposed method with i) the random sampling (denoted as random) that selects the samples randomly at each RF iteration, and ii) a double criteria AL (DCAL) method that considers the uncertainty and diversity criteria [10] which was previously used for CBIR problems in RS and thus is the most suitable reference for the proposed method. In the DCAL method, the uncertainty of images is evaluated by the MS strategy similarly to the proposed TCAL method. However, the diversity of images is assessed by simply estimating the distances of the most uncertain images in the image feature space and selecting the images that are most distant to each other. Moreover, we also compared the results of the proposed method that are obtained using the HI kernel with those obtained with the Radial Basis Function (RBF) kernel in order to evaluate the effectiveness of the HI kernel for the considered CBIR problems.

We carried out the experiments by adding h=5 samples at each iteration of AL and fixing the value of m (which is the number of images selected at the first step of the TCAL and DCAL methods) as 20 (i.e., 4h). We would like to point out the importance of the choice of the h value in

the design of the CBIR system to drive RF, as it affects the number of iterations necessary to reach convergence and thus both the performance and the cost of the retrieval system. In general, considering that time consuming image labeling is required at each AL iteration, the selection of high values for *h* may require high annotation time. Thus, it is much more effective to adopt small values for *h* and to carry out the required number of iterations for annotations until the user reach the desired retrieval performance.

According to the studied CBIR literature, results of each method are provided as: 1) learning rate graphs (which show the average precision on 30 trials versus the number of RF iteration), 2) standard deviation of precisions obtained on 30 trials; and 3) average precision-recall graphs where precision is plotted as a function of recall (which are obtained when the number of RF iterations is fixed). Precision is the fraction of retrieved images that are relevant (i.e., it measures ability to retrieve top-ranked images that are mostly relevant), and is obtained as the ratio between the number of relevant images retrieved and the number of all retrieved images. In the experiments, the most relevant 20 images are retrieved and the precision performance is evaluated on the top-20 retrieved images. Recall is the fraction of relevant images that are retrieved (i.e., it assesses the ability of the retrieval system to find all of the relevant images in the archive) and is obtained as the ratio between the number of relevant images that are retrieved and the number of all relevant images in the archive [24]. Note that for CBIR problems, precision-recall graphs are very useful in order to evaluate the retrieval performance on the top-ranked images (which is very common from a user point of view), as they assess retrieval performance at each point of ranking [24].

IV. Experimental Results

A. Analysis of the Effect of Histogram Intersection Kernel in CBIR

In the first set of trials, we analyze the effectiveness of the HI and RBF kernels in the framework of the considered CBIR problem. Fig. 5 shows the average precision versus the number of RF iteration obtained when the SVM classification and the proposed TCAL are implemented by

considering the HI and the RBF kernels for the retrieval of forest and agriculture images. Note that in the experiments the spread of the RBF kernel parameter is chosen performing a grid-search model selection, whereas the HI kernel has the advantage to be parameter free. By analyzing the figure, one can observe that HI kernel is much more effective than the RBF kernel for both categories. In other words, the results obtained using HI kernel achieves the highest precision at all the RF iterations. For example, the use of HI kernel yields a precision of 81.16% at the first RF iteration, whereas that of RBF kernel provides only 72% at the same RF round for the forest category (see Fig. 5a). Note that the differences on the precisions at the same RF round are very high at all RF rounds. These results clearly show the effectiveness of the HI kernel for comparing histogram-features in CBIR problems. On the basis of these results in the next of the paper we only report the performance provided by the proposed approach with the use of the HI kernel.

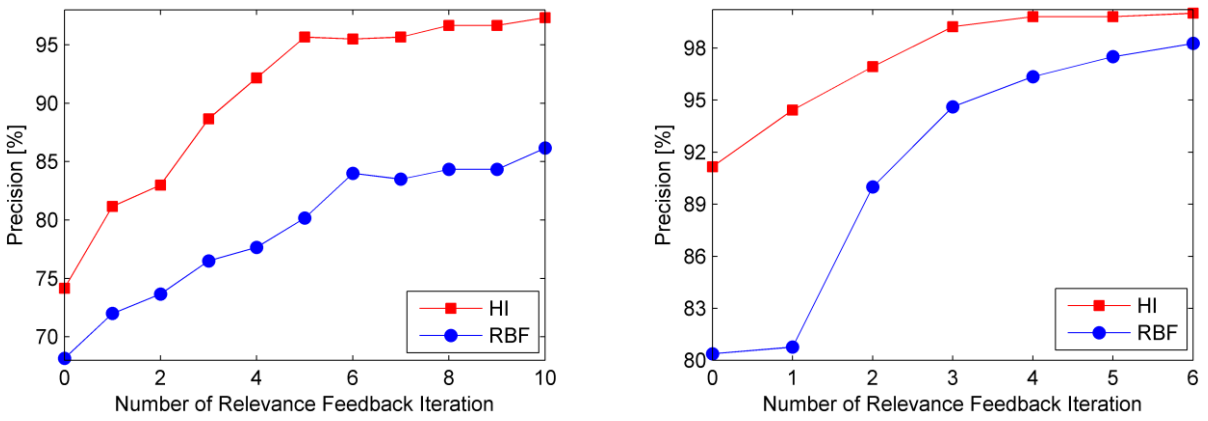


Fig. 5. Average precision versus number of RF iterations obtained by the proposed TCAL method when the HI and the RBF kernels are used for the retrieval of (a) forest and (b) agriculture images (when the top-20 images are retrieved).

B. Retrieval of Forest Images

In the second set of trials, we assess the effectiveness of the proposed TCAL technique in the retrieval of forest images. At first, in order to show the effectiveness of our choice on the selection of m (the number of uncertain samples that are being clustered) we carried out an analysis of the performances of TCAL and DCAL techniques varying its value as m=20, 50, 100. Fig. 6 shows the

precision versus the number of relevant feedback iteration obtained by the proposed TCAL. From the figure one can observe that small *m* values result in higher precision compared to that obtained selecting higher *m* values. This is particularly true at the later AL iterations. Note that at the initial iterations the results obtained with different *m* values are similar. This is due to the fact that at the initial iterations all the included images may significantly improve the performance since the set of annotated images is poor. However, at the later iterations it is important to select the most ambiguous ones and then assess the diversity and density on the selected images. Another interesting observation is that when using small *m* values, convergence is achieved with less annotated images than when using large values. Note that a similar behavior is also obtained by the DCAL method (we do not report the results for space constraints). On the basis of this analysis, we present the results obtained when *m*=20 in the rest of the paper both for the TCAL and DCAL methods.

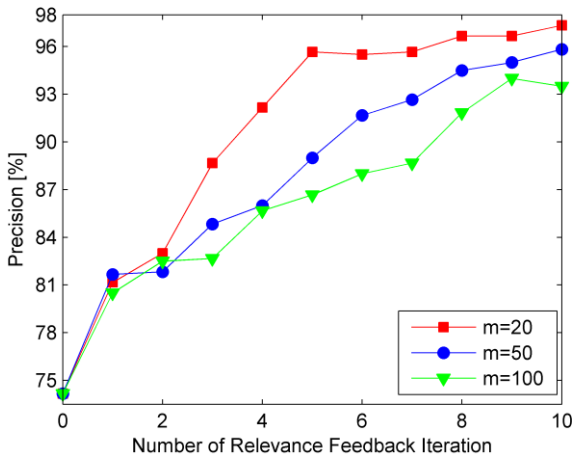
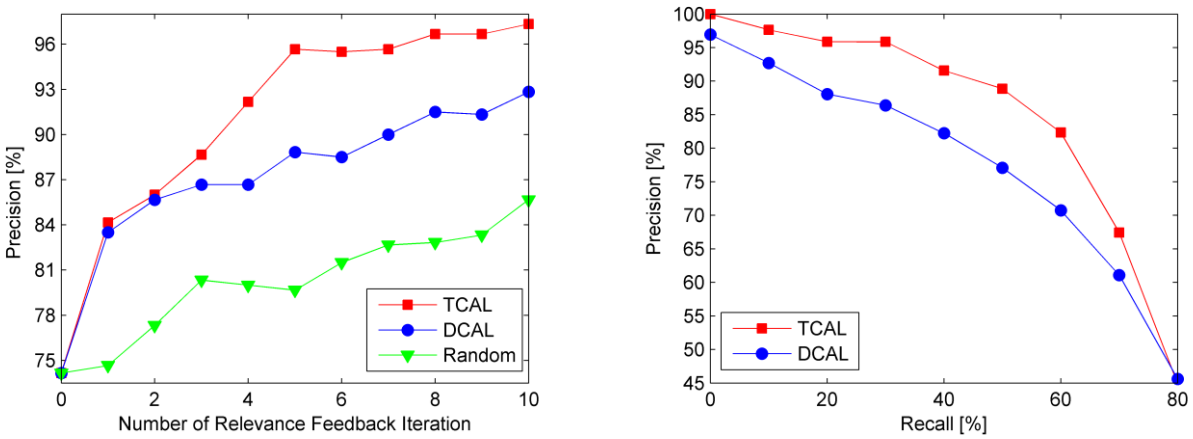


Fig. 6. Forest image retrieval: average precision versus number of RF iterations obtained by the proposed TCAL method, varying the value of *m* when the top-20 images are retrieved.

Fig. 7.a shows the behavior of the average (on 30 trials) precision versus the number of annotated images obtained when the top-20 images from archive are retrieved. In the figure, we compare the effectiveness of the proposed TCAL method with those of the random sampling and of the DCAL method presented in [10]. From the figure, one can see that the proposed TCAL

method leads to the highest precision at most of the RF iterations and significantly outperforms both the DCAL and the random sampling. Only at early iterations the TCAL provides similar precision to the DCAL, whereas it considerably increases the retrieval performance at the later iterations. As an example, the proposed TCAL achieves an average precision improvement of 7% over the DCAL and of 16% with respect to random sampling after the 5-th RF round when the number of total annotated images is 31 (see Fig. 7a). Both the TCAL and the DCAL methods achieve higher accuracies than the random sampling. Moreover, the TCAL method provides the same precision achieved by the DCAL with a smaller number of annotated images. For example, the TCAL method obtains a precision of 88.66% with 21 annotated images (i.e., at the 3-*rd* RF round), whereas the DCAL reaches similar precision with around 41 annotated images (i.e., at the 7-*th* RF round) (see Fig. 7a). These results show that selecting both uncertain and diverse unannotated images in the high density regions of the image feature space is very important since they are statistically very representative of the underlying image distribution and can significantly reduce the number of annotation required to achieve a given accuracy. Fig. 7.b shows the behavior of the precision-recall graphs obtained by the proposed TCAL and the DCAL methods after the 8-*th* RF iteration (i.e., annotated images is 46). By analyzing the figure one can see that the precision to recall ratio is improved by the TCAL with respect to DCAL method. In other words, the TCAL method always performs better than DCAL. Table 1 reports the mean and standard deviation of precision obtained on thirty trials versus different RF iteration numbers (and thus different training data size) for the TCAL, the DCAL and the random sampling. From the table we can observe that the precision obtained with the proposed TCAL are generally both higher and more stable (i.e., with lower standard deviation over the thirty trials) than those yielded by the DCAL and the random sampling. All these results point out the robustness and effectiveness of the proposed TCAL method to biased and imbalanced initial training sets in the archive.



(a) (b)

Fig. 7. Forest image retrieval: (a) average precision versus number of RF iterations obtained by the TCAL, the DCAL and the Random sampling when the top-20 images are retrieved, and (b) average precision versus recall graphs obtained by the proposed TCAL and the DCAL methods obtained after the 8-*th* RF iteration (when 46 images are annotated).

Table 1. Average and Standard Deviation (Std) of precisions obtained on 30 trials when top-20 images are retrieved at the 5-*th*, 8-*th* and 10-*th* RF rounds of the TCAL, the DCAL and the random sampling (Random).

Method	RF Iteration #5		RF Iteration #8		RF Iteration #10	
	Precision		Precision		Precision	
	Average	Std	Average	Std	Average	Std
TCAL	**95.66%**	**4**	**96.66%**	**4**	**97.33%**	**2**
DCAL	88.33%	9	91.50%	9	92.83%	9
Random	79.66%	21	82.83%	20	85.66%	18

Fig. 8 (which is related to one of the trials) shows a query image (see Fig. 8a) and the corresponding retrieved images obtained by the proposed TCAL and the DCAL methods when the number of annotated images is 51 (i.e., at the 9-*th* RF round). In the figure, the retrieval order of each image is reported. From the results one can see that most of the images retrieved by the proposed TCAL method belong to the forest category (see Fig. 8b) and are very similar to the query image. The 35-*th* retrieved image is belonging to river category in the archive; however it partially contains the forest class too. In this case the precision is 95%. On the contrary, the DCAL method returns irrelevant images also in the initial retrieval orders and provides a precision of 85%. For example, in Fig. 8 the 20-*th* retrieved image by the proposed TCAL is found relevant to

the query image (i.e., from the forest category), whereas the image retrieved at the same order by the DCAL method is irrelevant to the query image (i.e., from golf course). We also analyze the average precision-recall graphs obtained by fixing the number of AL iteration (and thus the number of annotated images). As we can see from the results, compared with the state-of-the-art DCAL method [10], our technique can effectively avoid problems caused by insufficient number of annotated samples in RF and thus is more effective for RS image retrieval.

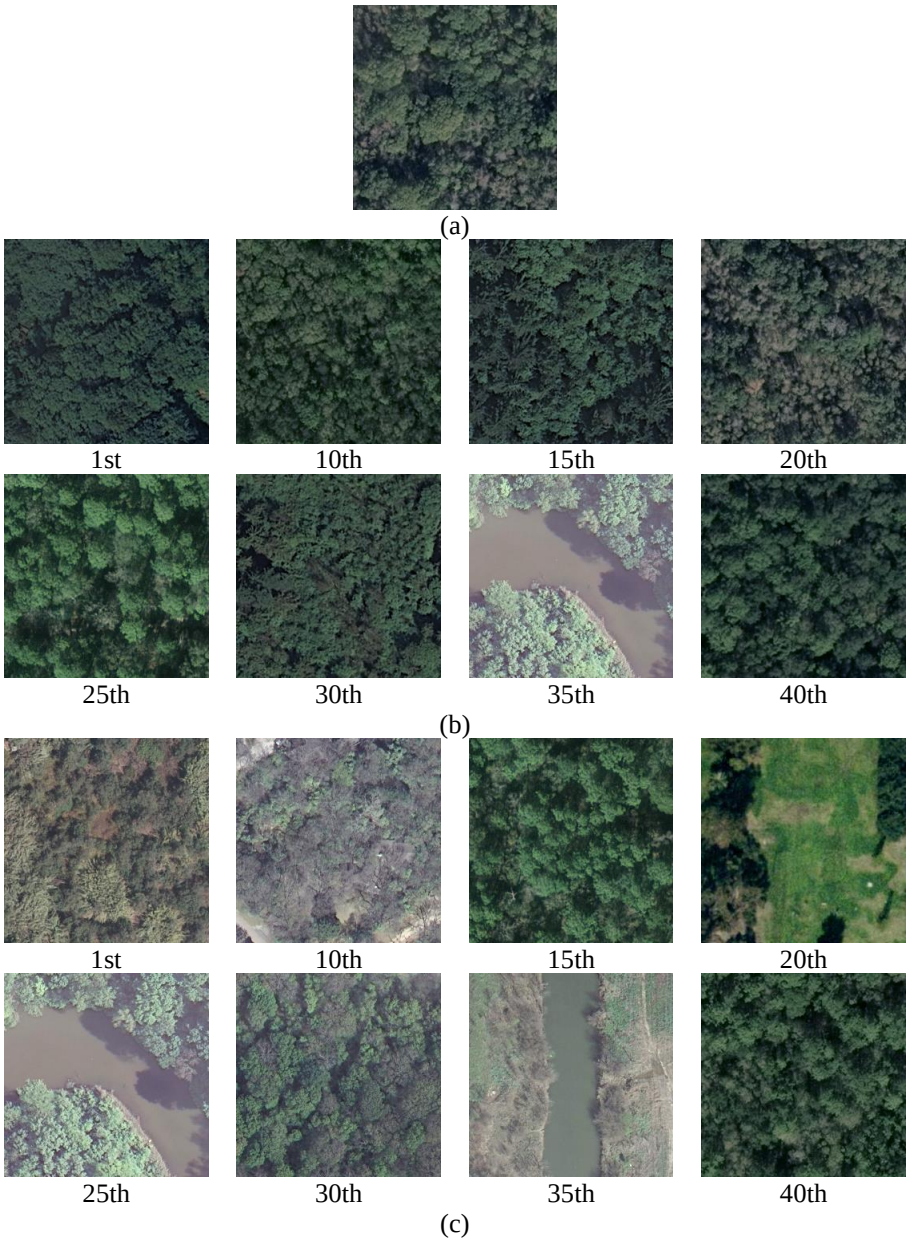
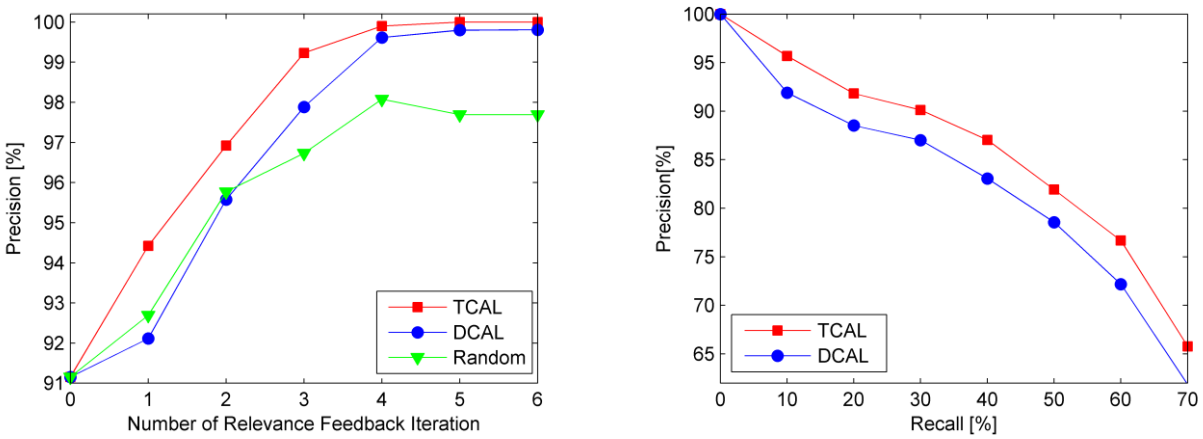


Fig. 8. Forest image retrieval: (a) query image; (b) example of retrieved images by the proposed TCAL (for which the retrieval accuracy is 95%); (c) example of retrieved images by the DCAL (for which the retrieval accuracy is 85%) in the case of retrieving top-20 images from the archive.

C. Retrieval of Agriculture Images

In the third set of trials, we assess the effectiveness of the proposed TCAL method when query images extracted from the agriculture category are considered. Fig. 9a shows the average (on 30 trials) precision versus the number of RF iteration when the top-20 images are retrieved. In the figure, we compare the effectiveness of proposed TCAL method with the random sampling and the DCAL [10]. By analyzing the figure, one can observe that the TCAL method again provides highest precision at most of the iterations. As an example, the TCAL technique achieves a precision of 94.42% at the 1-*st* RF iteration, whereas the DCAL and the random sampling provide precisions of 92.11% and 92.69%, respectively, at the same RF round when the top-20 images are retrieved (see Fig. 9a). From another viewpoint, the proposed TCAL method can provide the same precision obtained by the DCAL and the random sampling with a smaller number of annotated images (i.e., within less RF rounds). It is worth noting that random sampling requires much more image annotations to reach a similar accuracy. Fig. 9b shows the precision-recall graphs of the proposed TCAL and the DCAL methods obtained after the 1-*st* RF iteration (and thus the number of training samples is 11). From the figure one can again observe that the TCAL method outperforms the DCAL technique in terms of precision to recall ratio. Table 2 shows the average and standard deviation of precisions obtained on 30 trials versus different RF iteration numbers for the TCAL, the DCAL and random sampling. By analyzing the table we can again see that the average precision obtained with the proposed TCAL are higher and has a lower standard deviation (over the 30 trials) than those yielded by both the DCAL and the random sampling. Note that for this category the problem is simpler than for the forest category and thus the difference in results among the methods is reduced.



(a) (b)

Fig. 9. Agriculture image retrieval: (a) average precision versus number of RF iterations obtained by the TCAL, the DCAL and the Random sampling when the top-20 images are retrieved, and (b) average precision versus recall graphs obtained by the proposed TCAL and the DCAL methods obtained after the 1-*st* RF iteration (when the number of training samples is 11).

Table 2. Average and Standard Deviation (Std) of precisions obtained on 30 trials when top-20 images are retrieved at the 1-*st*, 2-*nd* and 5-*th* RF rounds of the TCAL, the DCAL and the random sampling (Random).

Method	RF Iteration #1		RF Iteration #2		RF Iteration #3	
	Precision		Precision		Precision	
	Average	Std	Average	Std	Average	Std
TCAL	**94.42%**	**13**	**96.92%**	**8**	**99.23%**	**2**
DCAL	92.11%	16	95.47%	10	97.88%	6
Random	92.69%	16	95.76%	12	96.73%	12

Fig. 10 shows a single trial of retrieval results with the corresponding query image (see Fig. 10a) and images retrieved by the proposed TCAL and the DCAL methods in the case of 21 annotated images (i.e., at the 3-th RF round). The retrieval order of each image is given below the related image. From the results one can see that all the images retrieved by the proposed TCAL method belong to the agriculture category (see Fig. 10b) and the precision on the top-20 retrieved images is 100%. On the contrary, the images retrieved by the DCAL method are not always related to the agriculture class. As an example, the 35-*th* and 40-*th* images retrieved by the DCAL method are associated with the forest and freeway categories in the archive, respectively (see Fig. 10c).In this case, the precision on top-20 retrieved images is 90% with the DCAL method.

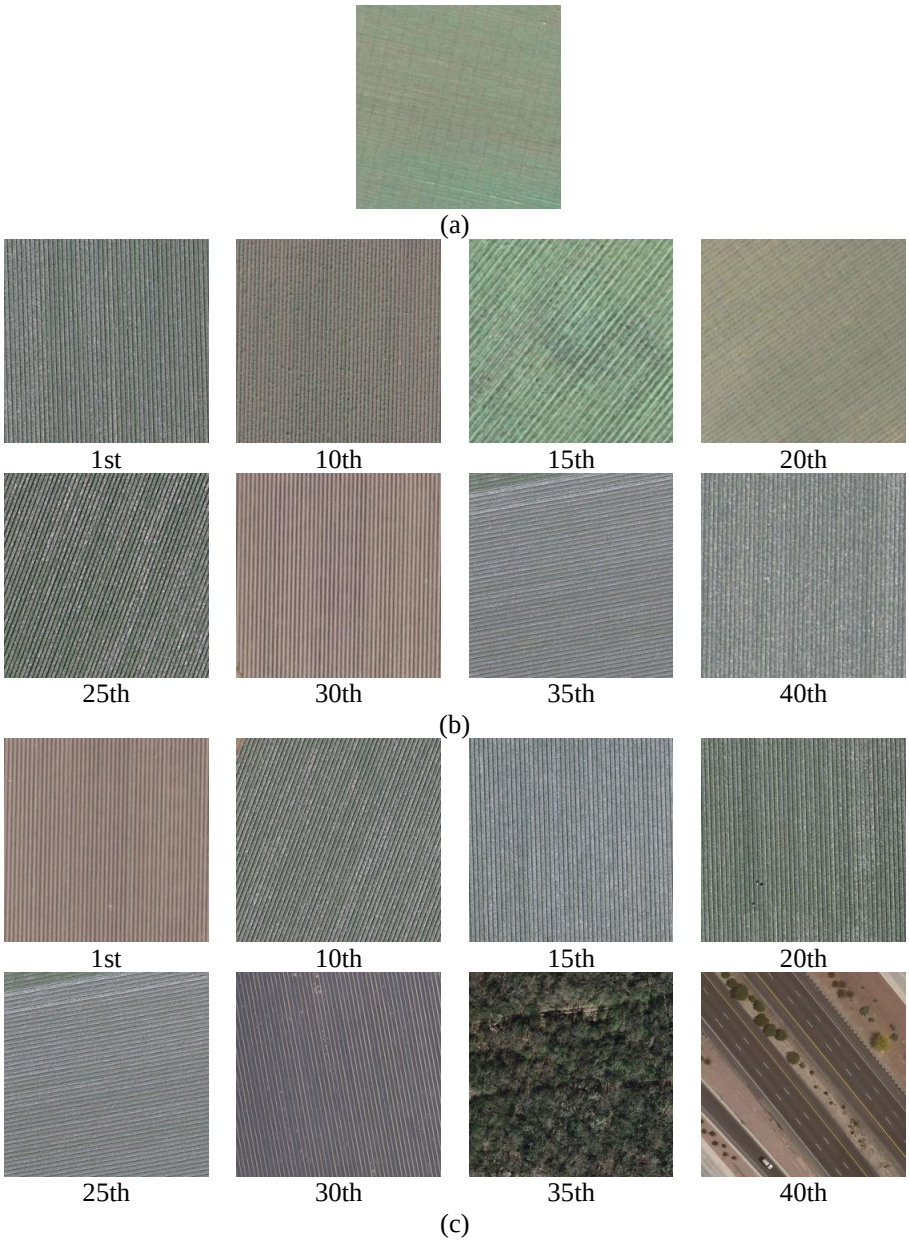


Fig. 10. Agriculture image retrieval: (a) query image; (b) images retrieved by the proposed TCAL (for which the retrieval accuracy is 100%); (c) images retrieved by the DCAL (for which the retrieval accuracy is 90%) in the case of retrieving top-20 images from the archive.

D. Retrieval of Images from all Categories

This subsection summarizes the results obtained by extracting the query images from all 21 categories. Accordingly, Fig. 11 shows average precisions in the top-20 retrieved results after the 10-*th* RF iteration for each category independently from other. The results are obtained randomly selecting 30 query images from each category and then averaging the results. As shown in Fig. 11,

the performance of the TCAL, the DCAL and the random sampling varies with different categories, whereas the proposed TCAL method always results in the highest precision for all the categories. In greater details, for difficult categories the average precision obtained by the proposed TCAL is much higher than those of the DCAL and random sampling (e.g., categories of Baseball diamond (11) and Storage tanks(20)). This clearly shows the ability of the proposed method to overcome the problems related difficult categories. For very easy categories, both the TCAL and the DCAL can perform well (e.g., categories of Harbor (6), Chaparral (12) and Parking lot (18)), whereas the random sampling shows lower performance. In general, compared with random sampling and DCAL, the proposed TCAL can perform much better for all of the 21 categories, indicating clearly the effectiveness of the proposed TCAL method for CBIR problems.

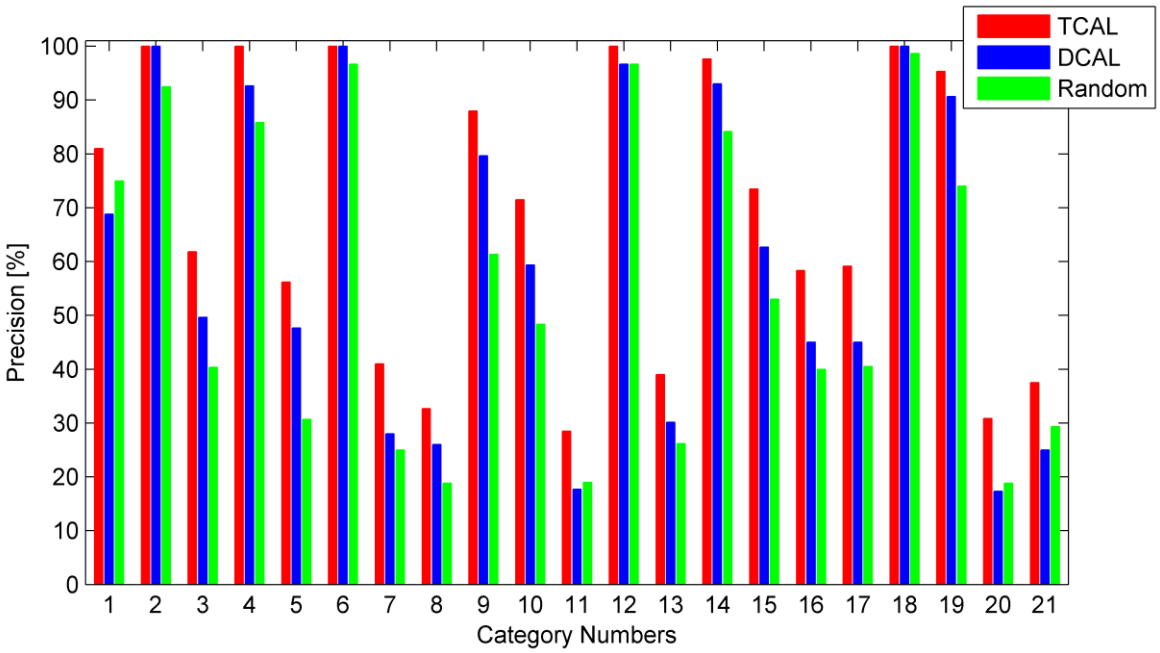


Fig. 11. Average precisions (on 30 trials per category, resulting 630 trials in total) obtained for different categories obtained by the TCAL, the DCAL and the Random sampling after the 10-*th* RF round when the top-20 images are retrieved (1:Beach; 2:Agriculture; 3:Buildings; 4:Forest; 5:River; 6:Harbor; 7:Dense residential; 8:Sparse residential; 9-Freeway; 10: Airplane; 11-Baseball diamond;12-Chaparral;13-Golf course;14-Mobile home park;15-Intersection;16-Medium residential;17-Overpass;18-Parking lot;19-Runway;20-Storage tanks;21-Tennis court).

Fig. 12 shows the average precision (on 30 trials per category, resulting 630 trials in total) versus number of RF iteration when the top-20 images are retrieved. From the figure, one can see

that general comments given for the categories of i) forest and ii) agriculture are confirmed for the whole archive on the 21 categories. In summary: 1) the TCAL method again results in the highest precision at most of the iterations with respect to the DCAL and random sampling; and 2) the proposed TCAL method can provide the same precision obtained by the DCAL and the random sampling with less RF rounds (i.e., with a smaller number of annotated images).

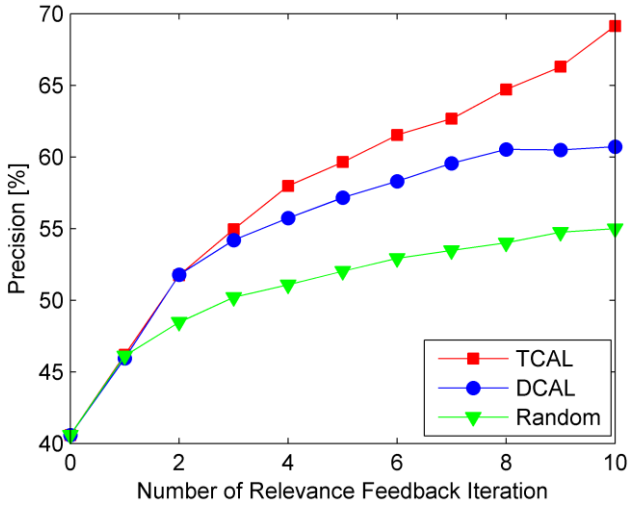


Fig. 12. Average precisions (on 30 trials per category, resulting 630 trials in total) versus the number of RF iterations obtained by the TCAL, the DCAL and the random sampling when the top-20 images are retrieved.

V. CONCLUSION

In this paper we have introduced a novel active learning (AL) method to drive relevance feedback in CBIR for the identification of effective images to annotate and to include in the training set. The proposed AL method selects both informative and representative unlabeled images to be included in the training set at each RF round by the joint evaluation of the uncertainty, diversity and density criteria. The uncertainty and diversity criteria aim to select most informative images, whereas the density criterion aims to select the most representative images in terms of prior distribution. In the proposed AL method the joint assessment of three criteria is accomplished based on a 2 steps technique. In the first step the most uncertain (i.e., informative/ambiguous) images are selected by using the well-known MS approach. In the second step the most diverse images among the

uncertain ones are selected from high density regions in the image feature space. In order to identify the highest density regions in the image feature space, a novel clustering based strategy has been introduced. The proposed AL method overcomes the limitations of previously presented AL methods in CBIR problems, which are due to:1) unbalanced training sets, and 2) biased initial training sets. Note that unlabeled images located in the high density regions of the image feature space are highly important for CBIR problems particularly when an unbalanced and biased training set is available. This is due to the fact that they are statistically very representative of the underlying image distribution, and thus the retrieval results on them affect much more the overall accuracy of the CBIR than those obtained on images within low density regions. Besides the overall system presented, the main novelties of the proposed AL method are: 1) the utilization of the prior term of the distributions based on the density of unlabeled images in the image feature space for driving the selection of images during RF rounds, and 2) the strategy to jointly evaluate the three criteria. Moreover, we introduce the use of histogram intersection kernel for CBIR problems (particularly in the context of the SVM classification and the proposed AL method) in RS.

The experimental performances of the proposed system were evaluated on an archive of 2100 images describing 21 different categories. The results show that the proposed AL method provides efficient image retrieval performance requiring less RF iterations and thus with less annotation effort compared to previously presented AL methods based on CBIR. We emphasize that these are very important advantages, because the main objective of AL in CBIR is to optimize the search with a minimum number of annotated images and thus with a minimum cost in annotating images.

It is worth emphasizing that given the growing amount of RS image archives, CBIR is becoming more and more important. One of the major challenges in CBIR is the semantic gap which can be reduced by RF driven by AL. Accordingly the proposed method is very promising as

it provides high retrieval accuracy with a small number of RF rounds. It is worth also nothing that the proposed AL method is independent from the considered feature extraction method, and therefore can be used with any feature extraction technique presented in the literature.

As a future development of this work, we plan to extend the validation of the proposed AL method to larger data sets and to use the proposed AL technique to drive the RF in image time series retrieval problems.

ACKNOWLEDGEMENT

This work has been partially supported by the Product Feature extraction and Analysis (PFA) project within the framework of the European Space Agency Long Term Data Preservation Program. The authors would like to thank Dr. Shawn Newsam for proving the remote sensing image archive used in the experiments.

REFERENCES

[1] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349-1380, Dec. 2000.

[2] R. Datta, D. Joshi, J. Li, and J.-Z. Wang, "*Image retrieval: ideas, influences, and trends of the new Age*," Journal ACM Computing Surveys, vol. 40, no. 2, pp. 1-60, 2008.

[3] P. Hong, Q.Tian, and T.S.Huang, "Incorporate support vector machines to content-based image retrieval with relevant feedback", *IEEE International Conference on Image Processing. Vancouver*, Canada, pp. 750-753, 2000.

[4] X. S. Zhou, T. S. Huan, "Relevance feedback in image retrieval: A comprehensive review", *Multimedia Systems*, vol. 8, pp. 536–544, 2003.

[5] Q. Bao and P. Guo, "Comparative studies on similarity measures for remote sensing image retrieval," *in Proc. IEEE Int. Conf. Syst., Man Cybern.*, The Hague, Netherlands, 2004, pp. 1112–1116.

[6] T. Bretschneider, R. Cavet, and O. Kao, “Retrieval of remotely sensed imagery using spectral information content,” *in Proc. IEEE Int. Geosci. Remote Sens. Symp*., Toronto, Canada, pp. 2253–2255, 2002.

[7] T. Bretschneider and O. Kao, “A retrieval system for remotely sensed imagery,” *in Proc. Int. Conf. Imag. Sci., Syst., Technol.*, 2002, Las Vegas, Nevada, USA, pp. 439–445.

[8] G. Scott, M. Klaric, C. Davis, and C.-R. Shyu, “Entropy-balanced bitmap tree for shape-based object retrieval from large-scale satellite imagery databases,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 5, pp. 1603–1616, May 2011.

[9] A. Ma and I. K. Sethi, “Local shape association based retrieval of infrared satellite images,” *in Proc. IEEE Int. Symp. Multimedia*, Irvine, CA, USA, 2005, pp. 551–557.

[10] M. Ferecatu and N. Boujemaa, “Interactive remote-sensing image retrieval using active relevance feedback,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 4, pp. 818–826, Apr. 2007.

[11] Y. Li and T. Bretschneider, “Semantics-based satellite image retrieval using low-level features,” *in Proc. IEEE Int. Geosci. Remote Sens. Symp*., Anchorage, AK, USA, 2004, vol. 7, pp. 4406–4409.

[12] Y. Hongyu, L. Bicheng, and C. Wen, “Remote sensing imagery retrieval based-on Gabor texture feature classification,” *in Proc. Int. Conf. Signal Process.*, Troia, Turkey, 2004, pp. 733–736.

[13] B. S. Manjunath and W. Y. Ma, “Texture features for browsing and retrieval of image data,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 837–842, Aug. 1996.

[14] S. Newsam, L.Wang, S. Bhagavathy, and B. S.Manjunath, “Using texture to analyze and manage large collections of remote sensed image and video data,” *J. Appl. Opt*., vol. 43, no. 2, pp. 210–217, Jan. 2004.

[15] A. Samal, S. Bhatia, P. Vadlamani, and D. Marx, “Searching satellite imagery with integrated measures,” *Pattern Recognit*., vol. 42, no. 11, pp. 2502–2513, Nov. 2009.

[16] S. Newsam and C. Kamath, “Retrieval using texture features in high resolution multi-spectral satellite imagery,” in Proc. SPIE Defense Security Symp., Data Mining Knowl. Discov.: Theory, Tools, Technol. VI, Orlando, FL, USA, 2004, pp. 21–32.

[17] Y. Yang, and S. Newsam, “Geographic image retrieval using local invariant features”, *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 2, pp. 818-832, Feb. 2013.

[18] L. Bruzzone, C. Persello, B. Demir, Active Learning Methods in Classification of Remote Sensing Images, in Signal and Image Processing for Remote Sensing, 2nd Edition, Ed: Prof. C.H. Chen, CRC Press – Taylor & Francis, Chapter 15, 2012, pp. 303-323.

[19] G. Schohn and D. Cohn, "Less is More: Active Learning with Support Vector Machines", *Proc. 17th Int'l Conf. Machine Learning*, Stanford, CA, USA, 2000, pp. 839-846.

[20] S. Tong and D. Koller, "Support Vector Machine Active Learning with Applications to Text Classification", Journal of Machine Learning Research, vol.2, pp. 45-66, 2001.

[21] K. Brinker, "Incorporating diversity in active learning with support vector machines", *Proceedings of the International Conference on Machine Learning*, Washington DC, 2003, pp. 59-66.

[22] R. Zhang, and A. I. Rudnicky, "A Large scale clustering scheme for kernel *k*-means", *IEEE International Conference on Pattern Recognition*, Quebec, Canada, 2002, pp. 289-292.

[23] B. Scholkopf, A. Smola, and K. R. Muller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*. vol. 10, no. 5, 1998, pp. 1299-1319.

[24] C. D. Manning, P.Raghavan, H. Schutze, Introduction to Information Retrieval, Cambridge University Press; 1 edition, 2008.

[25] A. Barla, F. Odone, A. Verri, "Histogram intersection kernel for image classification", *IEEE International Conference on Image Processing*, 2003, Barcelona, Catalonia, Spain, pp. III - 513-516.

[26] S. Maji, A.C. Berg, and J. Malik, "Classification using intersection kernel support vector machines is efficient," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Anchorage, AK, USA, 2008, pp.1-8.

[27] J. Wu, "A Fast dual method for HIK SVM learning", *11th European Conference on Computer Vision*, Heraklion, Crete, Greece, 2010, pp.552-565.