Osmam Emre Dai, Begüm Demir, Bülent Sankur, Lorenzo Bruzzone

A Novel System for Content-Based Retrieval of Single and Multi-Label High-Dimensional Remote Sensing Images

Dai, O. E., Demir, B., Sankur, B., & Bruzzone, L. (2018). A Novel System for Content-Based Retrieval of Single and Multi-Label High-Dimensional Remote Sensing Images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 11(7), 2473–2490.
https://doi.org/10.1109/jstars.2018.2832985

A Novel System for Content-based Retrieval of Single and Multi-Label High Dimensional Remote Sensing Images

Osman Emre Dai[1], Begüm Demir[2], Bülent Sankur[1] and Lorenzo Bruzzone[3]

[1]Electrical and Electronic Engineering Dept, Bogazici University, Istanbul, Turkey
[2]Faculty of Electrical Engineering and Computer Science, Technical University of Berlin, Berlin, Germany
[3]Dept. of Information Engineering and Computer Science, University of Trento, Trento, Italy

***Abstract*—This paper presents a novel content-based remote sensing (RS) image retrieval system that consists of: i) an image description method that characterizes both spatial and spectral information content of RS images; and ii) a supervised retrieval method that efficiently models and exploits the sparsity of RS image descriptors. The proposed image description method characterizes the spectral content by three different novel spectral descriptors that are: raw pixel values, simple Bag of Spectral Values descriptors and the extended Bag of Spectral Values descriptors. To model the spatial content of RS images we consider the well-known scale invariant feature transform-based bag of visual words approach. With the conjunction of the spatial and the spectral descriptors, RS image retrieval is achieved by a novel sparse reconstruction-based RS image retrieval method. The proposed method considers a novel measure of label likelihood in the framework of sparse reconstruction-based classifiers and generalizes the original sparse classifier to the case both single-label and multi-label RS image retrieval problems. Finally, to enhance retrieval performance, we introduce a strategy to exploit the sensitivity of the sparse reconstruction-based method to different dictionary words. Experimental results obtained on two benchmark archives show the effectiveness of the proposed system.**

***Index Terms*—Remote sensing, multi-label image retrieval,sparse reconstruction-based retrieval, spectral description, spatial description**

I. Introduction

With the continuous advances in satellite technology, the past decade has witnessed a tremendous growth in the availability of remote sensing (RS) image archives. Thus, content-based image retrieval (CBIR) has received extensive attention in the field of RS due to the necessity to search and retrieve in massive RS archives. CBIR aims to retrieve relevant images for a given query from very large image archives based on two main steps: 1) characterization of images by a set of features, and 2) retrieval of images that are similar to the query image.

To provide high performance CBIR results, image representation via discriminative and descriptive features is the fundamental step of a CBIR system. To this end, several methods have been introduced in the RS literature. Most of these methods are based on the extraction of low-level features such as intensity features [2], [3], [4], shape features [5], [6], [7], [8], and texture features [9], [10], [11], [12]. In recent years the global description of an image through the assembly of local features has proven to be very effective. In this paradigm, one generally uses the occurrence histogram of local features, mirroring the bag-of-words approach for text documents. Examples include the bag-of-morphological-words representation based on local morphological texture descriptors [15] and the bag-of-visual-words (BoVW) based on the scale invariant feature transform (SIFT) algorithm [16]. Another feature that has been studied is the local binary pattern (LBP) [13]. This feature describes local texture by comparing the value of a central pixel with those of surrounding pixels, and converting the sequence of comparisons into an integer. Hashing methods, which map the raw high-dimensional features into binary codes in a low-dimensional Hamming space, have recently received great attention in RS [14]. While these description algorithms were originally designed for grayscale images, they have been extended to multi-band (e.g. RGB, multispectral) images by the simple strategy of concatenating the descriptors that have been separately extracted from each band [17].

Despite the success of the descriptors in capturing spatial information for monochromatic images (which is the complete information content of single-band images), any adaptation for high-dimensional RS images falls short of efficiently expressing the spectral information. The extension of algorithms like SIFT for multi-band images, (i.e., applying the feature description algorithm on each band before concatenating the resulting descriptors) has no special consideration for spectral content and any improvement in performance over the grayscale application is rather incidental. As an example, a plain blue scene of a lake and a plain yellow scene of a desert would be considered to be exactly the same by the SIFT algorithm, which is based on the gradient information. Specifically, in this case the local descriptor would be the null vector for each image band. Another example of the approach that concatenates the per-band descriptors is given in [34]. In this work, a separate histogram is formed for each spectral band within a given window, and then these are concatenated to form the window's local spectral descriptor. However this approach has similar shortcomings, since the straightforward concatenation of per band descriptors not only would lead to high dimensionality (and thus high computational complexity), but also they would be expected to be highly correlated (and thus redundant). Furthermore the complete information content of a high dimensional image sample (e.g., hyperspectral image pixel) might not be fully exploited by constructing separate histograms at each band, as it does not explicitly consider the joint distribution of sample values for each band. Thus, we consider it crucial to have a local descriptor designed specifically to exploit both spatial and spectral information efficiently in RS images. It should be noted that some studies, most notably in the computer vision literature, have introduced algorithms specifically for high-dimensional images. The Color Descriptor by Luke et al. [20] expresses RGB values as hue-saturation vectors and builds SIFT-like descriptors using these color vectors instead of the grayscale gradient vectors of the original SIFT algorithm. These descriptors, designed specifically in the hue-saturation space, cannot be directly used for RGB images. The 'vectorized patches' used in [33] take a complete multi-band RS window, essentially a multispectral prism, and uses it as the local descriptor corresponding to the patch in vector form. However this method is likely to suffer from high-dimensionality when applied to hyperspectral images or when large image windows are considered. Another approach is the Bag of Colors [21], which quantizes the color space and expresses features based on histograms of color values. To our best knowledge this method has not to been implemented in the study of multispectral/hyperspectral RS images before.

Once image descriptors are obtained, one can proceed with the retrieval task. For example, one can use the k-nearest neighbor (k-nn) algorithm, which computes the similarity between the query image and all archive images to find the k images most similar to the query. However, this does not always result in satisfactory query responses due to the semantic gap, that is, the discrepancy between the image descriptors and high-level semantic content, although semantic gap problems can be somewhat alleviated by the relevance feedback methods. In contrast with k-nn, which relies on a similarity metric independent from the problem at hand (i.e., it includes no prior training), typical RS image retrieval systems are based on supervised classifiers by modelling retrieval as a binary-classification problem: one class includes images relevant to the query image, and the other class consists of irrelevant images. To this end, the Support Vector Machine (SVM) classifier has become popular in the framework of CBIR problems in the RS community [23]. Studies implementing such supervised classifier-based systems are mostly trained on images annotated with single high-level land-use category labels, which are associated to the most significant content of the image [16]. In reality RS images typically possess multiple classes and thus can simultaneously be associated with different semantic labels (i.e., low-level land-cover class labels). Thus, supervised CBIR methods that properly exploit training images annotated by low-level land-cover class labels (i.e., multi-labels) are required. To address this problem, multi-label learning methods have been recently found very promising in the multimedia communities for multi-label image search and retrieval problems [24], [25], [26]. In [24] the classical k-nn approach was enhanced to achieve multi-label classification by including a statistical analysis, and allowing the use of the maximum a posteriori principle. In [25] the authors use an SVM classifier to perform one-versus-all binary classification for each category, which allows one to choose the single most relevant category among multiple others. An extensive study comparing various multi-label classification methods is given in [26]. These methods do not model and exploit the sparsity often present in high dimensional image features during the retrieval, and thus may provide low retrieval performance when the image descriptors are high dimensional and sparse.

To overcome these limitations, that is, high dimensionality and redundancy of image descriptors, and the need to better exploit both spatial and spectral information content, we propose a CBIR system that includes: 1) an image description method that incorporates both spectral and spatial information of RS images; and 2) a supervised retrieval method that is suitable to exploit training images annotated by a single high-level land-use category label or by multiple low-level land-cover class labels. For this purpose, two parallel description algorithms for RS images are executed to extract components of spatial and spectral information. For spectral description, we propose a histogram-based novel local descriptor that we call the bag of spectral values (BoSV). Local BoSV descriptors are extracted from the complete RS image and the histogram of these descriptors is used to express spectral content. For spatial description we first apply dimension reduction on an RS image followed by the per band SIFT-based BoVW formalism to efficiently express spatial features. The proposed supervised retrieval method aims to: 1) efficiently model and exploit the sparsity of features; and 2) be applicable for both single-label and multi-label RS image retrieval problems. To this end, our retrieval method uses the sparse coding of global image descriptors. This is inspired from the sparse reconstruction-based classifier (SRC) [27], which has been shown to be robust even in the case of limited training samples and large-sized, sparse image representations. Furthermore, we use a novel measure of label likelihood in the context of sparse reconstruction-based classifiers and generalize the original sparse classification approach to the case of multi-label images: For images annotated with multi-labels, we consider first each label individually, and estimate the likelihood of an image to contain that specific label independently of the remaining labels. This is done by checking the sparse reconstruction performance on a limited set of annotated training images. We also propose a strategy that weighs individual features according to their significance. All in all, the three novelties we bring in to the RS CBIR art are: 1) an image description method that models and exploits spatial and spectral information of RS images jointly; 2) adaptation of the sparse classifier to single-label and multi-label RS image retrieval problems; and 3) a strategy to improve retrieval performance by boosting the significance of relevant features. Our system has been briefly presented in [1] with limited experimental results. The current paper builds upon this work and contributes in several ways. We add two new local spectral descriptors to the spectral description algorithm and we extend the multi-label retrieval approach to cover applications on category-based single label archives. Furthermore several new experiments are run and their results commented.

The rest of the paper is organized as follows: Section II provides the problem formulation and describes the related works. Section III explains the proposed spatial-spectral image description method, while Section IV describes the proposed RS image retrieval method based on sparse reconstruction. Section V describes the experimental setting and benchmark archives on which the proposed system has been evaluated. Section V provides the results with a discussion and Section VI draws the conclusion of the work.

II. Preliminaries

A. Problem Formulation and General Notation

Let $\boldsymbol{X} = \{X_1, \ldots, X_R\}$ be an RS archive of R images where X_i is the i-th image. Given a query image X_q, the retrieval problem consists of determining a ranked list of images from $\boldsymbol{X}$ that are most similar to X_q. We also assume the availability of an annotated training set $\boldsymbol{X}_S = \{X_1^S, \ldots, X_M^S\}$ of M RS images. The annotation may be of two types: 1) it is in the form of high-level land-use categories such as residential area, farmland or forest (in this case each image is matched with a single such class); and 2) it is based on low-level land-cover classes, such as grass, asphalt or sand (this case, depending on the complexity of the scene, images may contain one or more of these labels.) In the latter case, the retrieval is based on some similarity score between the likelihoods of the labels of the query and archive images.

Concerning single-labeled annotation, we express the training set as a union of image sets, each characterized by one category: $\boldsymbol{X}_S = \boldsymbol{X}_1 \cup \boldsymbol{X}_2 \cup \ldots \cup \boldsymbol{X}_C$ where the c-th set $\boldsymbol{X}_c = \{X_{c_i}^S\}_{i=1}^{M_C}$ incorporates all training images belonging to the category $c \in \{1, \ldots, C\}$. Concerning multi-labeled annotation, we define a label set $\Phi = \{\phi_1, \ldots, \phi_{|\Phi|}\}$ of $|\Phi|$ labels; the labels of an archive image X_i^S consist of a subset of the label set: $\Phi_i \subset \Phi$. The label information of X_i^S can also be expressed by a binary vector $\boldsymbol{L}(X_i^S) \in \{0,1\}^{|\Phi|}$ where the value at the c-th position $\boldsymbol{L}^c(X_i^S)$ corresponds to the presence or absence of the label and vice versa: ϕ_c, i.e., $\boldsymbol{L}^c(X_i^S) = 1 \iff \phi_c \in \Phi_i$.

The content of an image is expressed using the Bag of Visual Words approach where one considers the occurrence statistics of certain well-defined visual words $\mathbf{W} = \{w_k\}_{k=1}^K$. For any image X_i we express this "bag" as a histogram of occurrences $\boldsymbol{f}_i = [f_i^1, \ldots, f_i^K]^\top$, where f_i^k corresponds to the frequency of occurrence for word w_k. In addition to BoVW, the proposed system makes use of a variant of the SRC algorithm. We briefly present these two concepts before giving the details of our system.

A list of all mathematical symbols used throughout the paper is given in Appendix B in Table VII.

B. Bag of Visual Words Approach

The Bag-of-Visual-Words approach is a well established paradigm in image description where an image is characterized by the occurrences of visual pieces, i.e., words that it contains. It has been observed that for the RS images the location of visual words and the structure of the image is often not crucial, though the construction of a representative set of visual words remains critical [28]. The number or frequency of instances of words is sufficient to deduce information about an image's content and therefore it becomes possible to represent a document as an unordered set (or "bag") of words.

Various algorithms, such as the Scale Invariant Feature Transform (SIFT) [29], are commonly used to extract the local descriptions of images. The local descriptors can either be extracted from salient patches, i.e., at interest points found by a detection algorithm, or based on a dense visiting schedule from neighborhoods of pixels. A visual dictionary is then built using a clustering algorithm and a training set of image features. Given a visual dictionary, the content of any image can be expressed as a histogram of the visual words. The histogram of occurrence frequencies of the dictionary elements becomes the BoVW representation of the image. This approach has proven to be a very useful method for classification and retrieval tasks of grayscale images whereas its extension to multispectral and hyperspectral RS images is not straightforward and remains as a challenge.

C. Sparse Reconstruction-based Classification

Sparse representation-based classification (SRC) has been popularized with the work of Wright et al. [27], where for any image the system computes a vector of sparse coding coefficients in terms of other class specific similar images, i.e., the training set. SRC classification decides on a class based on the sparse reconstruction residuals from training images of the respective classes. This idea of sparse reconstruction can be generalized to images represented by their global descriptors, such as their BoVW representations.

Let $\boldsymbol{f_i}$ be the global image descriptor to represent image X_i and let $\boldsymbol{F}_c = \left\{(\boldsymbol{f}_{c_i})^S\right\}_{i=1}^{M_c}$ be the sets of archive global image descriptors belonging to different categories $c \in \{1, ..., C\}$. The size of the training set is $M = M_1 + \ldots + M_C$. SRC method decides for a class based on the capability of the global descriptor $\boldsymbol{f_i}$ of an image X_i to be reconstructed as a linear combination of the global descriptors of training images $\boldsymbol{F}_c$. Let $\mathbf{D_c} = \left[\boldsymbol{f}_{c_1} \ldots \boldsymbol{f}_{c_{M_c}}\right]$ be the matrix formed of global descriptors of training images from class c and dictionary $\mathbf{D} = [\mathbf{D}_1 \ldots \mathbf{D}_C]$ be the matrix of all global descriptors. Sparse coding can be expressed as the following optimization problem:

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^M} \|\boldsymbol{f}_i - \mathbf{D}\boldsymbol{\alpha}\| \ \text{ such that } \ \psi(\boldsymbol{\alpha}) \leq \eta \tag{1}$$

where $\boldsymbol{\alpha}$ is the sparse code, $\boldsymbol{f_i}$ is the global descriptor to be sparse coded, $\mathbf{D}$ is the sparse coding dictionary, η is the sparsity threshold and function ψ is the sparsity-inducing penalty, which is a measure inversely related to the sparsity of the solution. In our application the ℓ_0-norm is used as the sparsity-inducing penalty. For this case, the problem in (1) can be expressed as choosing at most η columns of $\mathbf{D}$ enabling the linear reconstruction of $\boldsymbol{f}_i$ with minimum error. A solution to this problem can be computed using the greedy Orthogonal Matching Pursuit (OMP) algorithm [30].

The class decision is based on the comparison of the potential of the respective training sets to reconstruct the descriptor $\boldsymbol{f}_i$. For this purpose, consider the decomposition of a global descriptor in terms of sparse reconstructions:

$$\boldsymbol{f}_i \approx \mathbf{D}\boldsymbol{\alpha} = \mathbf{D}_1\boldsymbol{\alpha}_1 + \ldots + \mathbf{D}_C\boldsymbol{\alpha}_C \quad \text{where} \quad \boldsymbol{\alpha} = \left[\boldsymbol{\alpha}_1^\top \ldots \boldsymbol{\alpha}_C^\top\right]^\top \tag{2}$$

The residual $r(\boldsymbol{f}_i, c)$ corresponding to the category c is then expressed as:

$$r(\boldsymbol{f}_i, c) = \|\boldsymbol{f}_i - \mathbf{D}_c\boldsymbol{\alpha}_c\| \tag{3}$$

Then, the class label c_i to be assigned to the i-th image is identified as:

$$c_i = \arg\min_c r(\boldsymbol{f}_i, c)$$

With this approach one assigns the most likely class to any given image. However this method does not provide an explicit metric to be used for the inverse problem, which is more relevant to image retrieval for determining the most likely images to belong to a given class. Furthermore this approach is only valid in the case of high-level land-use category labels, where each image is associated with a single label. For archives where a single image can be related with multi-labels, such a category-based comparison can not be applied. We provide solutions to these problems in Section IV.

III. Proposed Spatial - Spectral Image Description Method

We propose an RS image description method to efficiently characterize both spatial and spectral information content of images. This is achieved by processing the spatial and spectral contents individually in two parallel pipelines, each one expressing the content information via a global descriptor according to its own BoVW step (see 1). In the proposed method, the local spectral descriptors are obtained through a novel description algorithm, while the spatial description relies on the well known local SIFT descriptors. The global spatial descriptor $\boldsymbol{f}_i^{spat}$ and the global spectral descriptor $\boldsymbol{f}_i^{spect}$ of the i-th image X_i obtained from these pipelines are then scaled by weights $(1 - \lambda_{spect})$ and λ_{spect}, respectively, and stacked together to form a global descriptor $\boldsymbol{f}_i$ (which we henceforth refer to as the hybrid descriptor) whose length equals the sum of the respective global descriptor lengths.

$$\boldsymbol{f}_i = \left[(1 - \lambda_{spect}) \times (\boldsymbol{f}_i^{spat})^\top, \lambda_{spect} \times (\boldsymbol{f}_i^{spect})^\top\right]^\top \tag{4}$$

The weight of each global descriptor (i.e. λ_{spect} for $\boldsymbol{f}_i^{spect}$ and the complement for $\boldsymbol{f}_i^{spat}$) corresponds to the significance of the information it provides. For example, an increase in the spectral resolution would be expected to augment the information in the spectral pipeline and thus should be accorded a proportionally higher value for λ_{spect}. The proposed methods used to extract spatial and spectral information extraction pipelines are given in the sequel.

A. Spatial Image Description

For spatial description of RS images, while any local spatial descriptor can be used in our framework, we have adopted the well-known SIFT-based BoVW approach [16]. Let the number of spectral bands in images from $\boldsymbol{X}$ be N. First, the principal component analysis (PCA) is performed on a randomly selected set of K-dimensional (multi-band) pixels from randomly selected images of archive $\boldsymbol{X}$. The number of principal components $n < N$ for spatial description is determined to satisfy a certain percentage of the total variance. Given the n-dimensional principal component vectors, the N-dimensional pixel samples are projected onto the n-dimensional subspace of chosen principal components. Thus the N-band X_i images are reduced to more compact n-band images $\widetilde{X}_i$ without significant loss of information, but enabling more efficient feature extraction. In this paper the dense SIFT algorithm is applied on the images $\widetilde{X}_i$, i.e., SIFT features are extracted on a uniform grid from the PCA projected images. Note that since dense SIFT disregards dominant orientations, it lacks the robustness of the original interest point-based SIFT algorithm against rotations. Therefore we implement a simple algorithm to compute the dominant orientation for each local descriptor and then re-orient it accordingly. Recall that SIFT vectors are formed from multiple histograms of gradient orientations corresponding to different sub-regions around a keypoint [29]. Thus each vector entry represents a histogram bin in some sub-region, corresponding to one of eight main orientations. Let us denote them by cardinal and intercardinal directions: E, NE, N, NW etc. We sum up values corresponding to each cardinal direction. For example, we sum the three histogram bin values, NW, N and NE, to obtain the dominance score of the northward orientation. In this summation we do not weight any direction in order not to mitigate the contribution of intercardinal directions, because the latter may be significant. Note that this implies that each intercardinal direction contributes with equal weight to its two adjacent cardinal directions. Once the dominant orientation for each descriptor is computed, we apply a linear transform on the vectors to *reorient* the local SIFT descriptors such that each of them has the same dominant orientation. This is done simply via a permutation of vector entries. For example, to reorient an eastward-orientated SIFT vector such that the dominant orientation becomes north, all histogram bins expressing the N orientation are replaced by those expressing E, NW by NE, W by N etc. This approach provides a certain degree of rotation-invariance. This is a simplified version of the approach presented in [29], where reorientations were not limited to only 4 directions. We do not follow this approach for two reasons: i) Our dense feature algorithm results in a much higher number of local descriptors than the sparse SIFT algorithm in [29], making complex descriptor reorientation methods impractical; in fact, the resulting computational load would be aggravated for higher spatial resolution images. ii) Non-cardinal reorientation requires interpolation on pixel values, which in the hyperspectral case might potentially result in unnatural spectral values and hence less meaningful local spectral descriptors. This issue would be especially significant

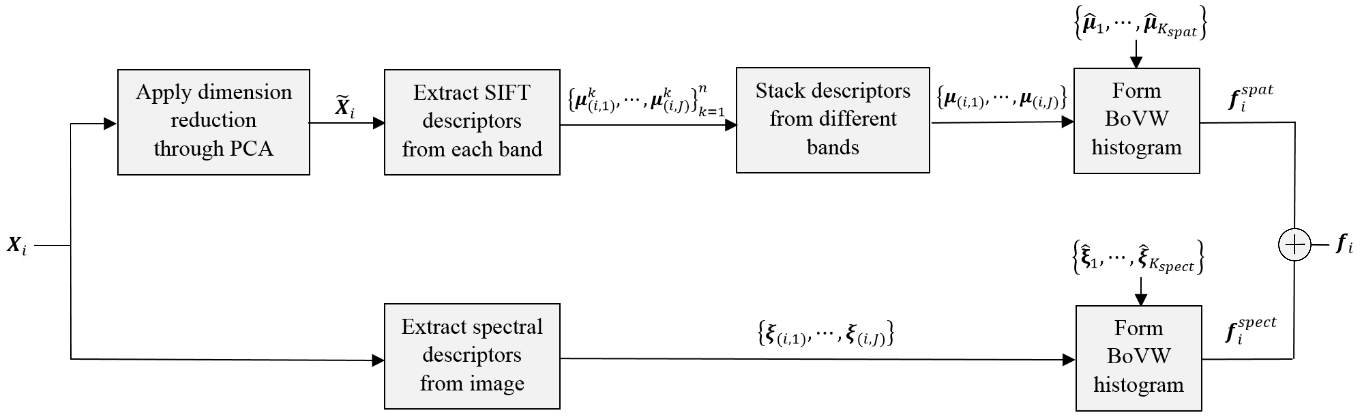


Fig. 1. Block diagram of the proposed spatial-spectral RS image description method with separate spatial feature extraction pipeline (above) and spectral feature extraction pipeline (below).

for low spatial resolution archives, since pixel values become less representative of their neighbors as distances increase, compromising the meaningfulness of interpolated values.

We apply the algorithm on each of the n bands to extract a set of J local SIFT descriptors for each band $\{\boldsymbol{\mu}^1_{(i,1)}, \ldots, \boldsymbol{\mu}^k_{(i,j)}, \ldots, \boldsymbol{\mu}^b_{(i,J)}\}^n_{b=1}$. In this expression, i denotes the index of the image, b denotes the index of the band from which the local descriptor is obtained, and j denotes the index of the pixel location from which the local descriptor is extracted. Note that all descriptors with the same subscript j belong to the same pixel location across all bands $b \in \{1, \ldots, n\}$. These descriptors are then concatenated to form local multi-band SIFT descriptors (which we refer to as local mSIFT descriptors) $\boldsymbol{\mu}_{(i,j)} = \left[(\boldsymbol{\mu}^1_{(i,j)})^\top, \ldots, (\boldsymbol{\mu}^n_{(i,j)})^\top\right]^\top$. A local SIFT descriptor of size 128 is obtained per each of the n PCA bands, hence the local mSIFT descriptors are of size $128 \times n$. The final expression (i.e. the global spatial descriptor) is obtained by applying the BoVW approach on the local mSIFT descriptors. A codebook containing K_{spat} descriptors $\{\widehat{\boldsymbol{\mu}}_1, \ldots, \widehat{\boldsymbol{\mu}}_{K_{spat}}\}$ is formed by randomly collecting local descriptors from the archive of images and clustering them into K_{spat} clusters. This set then forms the codebook of spatial descriptors (i.e. codebook of spatial visual words). Given this codebook, then local multi-band SIFT descriptors of an image are associated with the closest words from the $\{\widehat{\boldsymbol{\mu}}_1, \ldots, \widehat{\boldsymbol{\mu}}_{K_{spat}}\}$ set. The occurrence histogram of these words form a K_{spat}-dimensional histogram $\boldsymbol{f}_i^{spat}$ for any X_i, which is the global descriptor for the spatial content of the image. The steps of the spatial description pipeline are illustrated on the top section in Figure 1.

B. Spectral Image Description

For spectral description $\boldsymbol{f}_i^{spect}$, $i = 1, \ldots, R$ of RS images, a different BoVW approach is implemented using a specific spectral description algorithm in lieu of the SIFT algorithm. As shown in the lower branch on Figure 1, this specific algorithm takes the complete RS image X_i as input (without applying any dimension reduction) and produces a set local descriptors $\{\boldsymbol{\xi}_{(i,1)}, \ldots, \boldsymbol{\xi}_{(i,J)}\}$ extracted from a dense grid on the image (the detailed explanation of these local descriptors are given in the next paragraph). A K_{spect}-sized codebook of these local descriptors $\{\widehat{\boldsymbol{\xi}}_1, \ldots, \widehat{\boldsymbol{\xi}}_{K_{spect}}\}$ is formed by clustering from a sample set of N-dimensional pixels. Given an RS image X_i, the histogram $\boldsymbol{f}_i^{spect}$ of length K_{spect} serves as the global descriptor of spectral features. Three alternative local spectral descriptors are considered: raw pixel values (RP), simple Bag of Spectral Values (SBoSV) descriptors, and extended Bag of Spectral Values (EBoSV) descriptors.

1) Raw Pixel Description (RPD): Raw pixel description consists of a simple approach where the N-dimensional pixel values are directly taken as local spectral descriptors. This straightforward approach eliminates the need for dimension reduction or quantization in the local processing prior to the extraction of image descriptors. Note that although local descriptors are quantized before obtaining the global descriptor $\boldsymbol{f}_i^{spect}$ of X_i, the fact that local descriptors have not undergone any prior quantization could be expected to minimize information loss. The simplicity of the approach offers some advantage in terms of computational load. The shortcoming of this method is that it considers pixels as isolated entities, completely ignoring their adjacency and spatial memory.

2) Simple Bag of Spectral Values (SBoSV): SBoSV is defined as the histogram of spectral values within a window as shown in Figure 2.c. These local descriptors are extracted from overlapping patches from the RS image using a sliding window approach. We assume that a global codebook of spectral values has been computed beforehand using pixels randomly selected from the archive images. This approach enables for local descriptors that express the spectral composition within a neighborhood instead of only at a single pixel, and thus incorporates some local spatial information. Recall that the raw pixel description above completely ignores the content distribution within neighborhoods.

3) Extended Bag of Spectral Values (EBoSV): EBoSV is a enhancement on the SBoSV in that, in addition to describing the pixel content within a window, it gives some idea on how these pixels are distributed by considering five separate regions within the windows of interest. The EBoSV divides the window into multiple sub-windows corresponding to the upper,

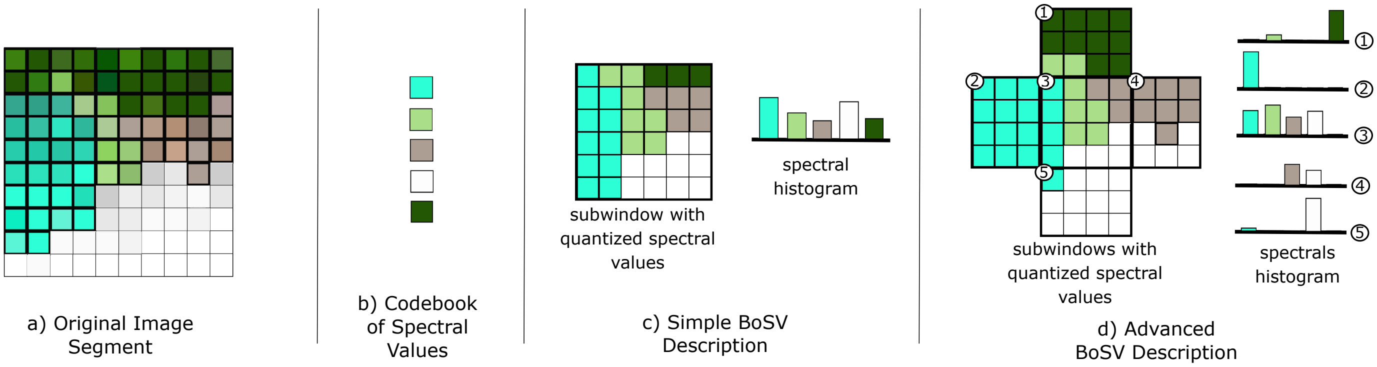


Fig. 2. Illustration of the proposed local spectral descriptors: a) an image patch prior to quantization, b) a spectral codebook of five visual words, c) the coded patch and its resulting histogram for SBoSV, d) the coded patch and its resulting histograms for EBoSV.

lower, left, right and central regions and forming a separate histogram for each of them before finally concatenating them all into a single local descriptor (see Figure 2.d). Considering that this descriptor is not invariant to orientations, we have implemented the same descriptor reorientation algorithm we use for SIFT descriptors to the EBoSV descriptors. EBoSV leads to more informative local spatial descriptors at the cost of more computations, since the descriptors are five times longer than those in SBoSV.

Note that, unlike the studies in [20], [33] all three local descriptors are applicable to RS images with an arbitrary number of bands to characterize image windows of any size. Also note that for both spatial and spectral description, while one can expect to end up with visual codebooks of large sizes, and hence very high dimensional global descriptors $\boldsymbol{f}_i$ depending on the inherent complexity of the archive; for individual images one can expect reasonably sparse global descriptors despite the large size of the codebook. This will be, in fact, an opportunity for the sparsity-based retrieval phase.

IV. Proposed Image Retrieval method based on Sparse Reconstruction

To efficiently exploit sparse high-dimensional spatial-spectral descriptors that we obtain in the image description phase of the proposed system, we propose two sparse reconstruction-based CBIR methods. The proposed methods are based on a novel measure of label-likelihood in the framework of sparse reconstruction-based classifiers and they generalize the original sparse classifier to the cases of both single-label and multi-label RS image retrieval. The next two subsections give details on both methods, and in the final subsection we introduce a strategy to assign weights to the visual words to further enhance retrieval performance.

A. Proposed Single Label RS Image Retrieval Method

The outline of our proposed single label RS image retrieval method has been given in Fig. 3. Let $\boldsymbol{f}_q$ be the global descriptor of the query image X_q known to belong to the high-level land-use category $c_q \in \{1, \ldots, C\}$. Let $\boldsymbol{F}_c = \left\{(\boldsymbol{f}_c)^S\right\}_{i=1}^{M_{c_i}}$ global descriptors of images from the training set $\boldsymbol{X}_c$ for each land-use category $c \in \{1, \ldots, C\}$, and $\boldsymbol{F} = \{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_R\}$ be

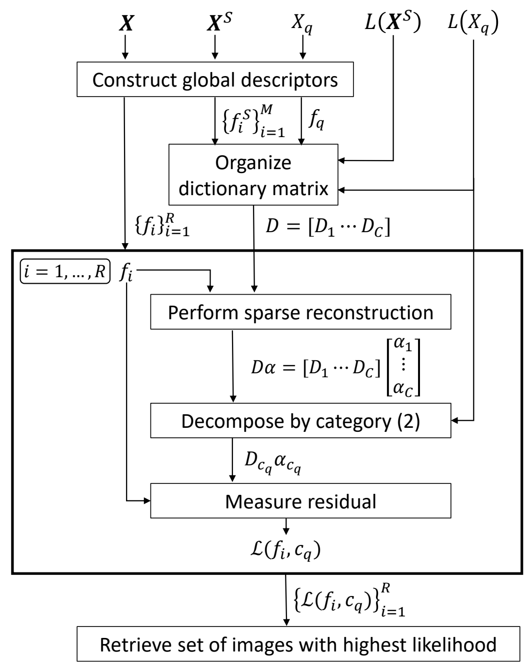


Fig. 3. Block diagram of the proposed single label retrieval method. Note that all label vectors $\mathbf{L}(X_i)$ are binary vectors with a single non-zero entry. Specifically for $\mathbf{L}(X_q)$ that entry is located in position c_q

the set of global descriptors of images from the unannotated archive $\boldsymbol{X}$. We seek the images from $\boldsymbol{X}$ whose global descriptors are most similar to the query descriptor $\boldsymbol{f}_q$, ideally retrieving images all belonging to the land-use category c_q of the query image X_q. This method requires the formation of category-specific dictionaries $\mathbf{D}_c = \left[\boldsymbol{f}_{c_1}, \ldots, \boldsymbol{f}_{c_{M_c}}\right]$(similar to the SRC classifier, see Section II.C). Given the set of these dictionaries, one first performs a sparse reconstruction of the global descriptor $\boldsymbol{f_i}$ of image X_i (a candidate for retrieval) based on the union of all dictionary sets $\mathbf{D} = [\mathbf{D}_1, \ldots, \mathbf{D}_C]$. Then sparse reconstruction residuals are calculated as in (3). The premise is that images belonging to a given land-use category have lower residual energy in that category, and thus a better representation, in comparison with the residual energies of images belonging to other land-use categories. However, for a given image, unrelated categories can sometimes achieve

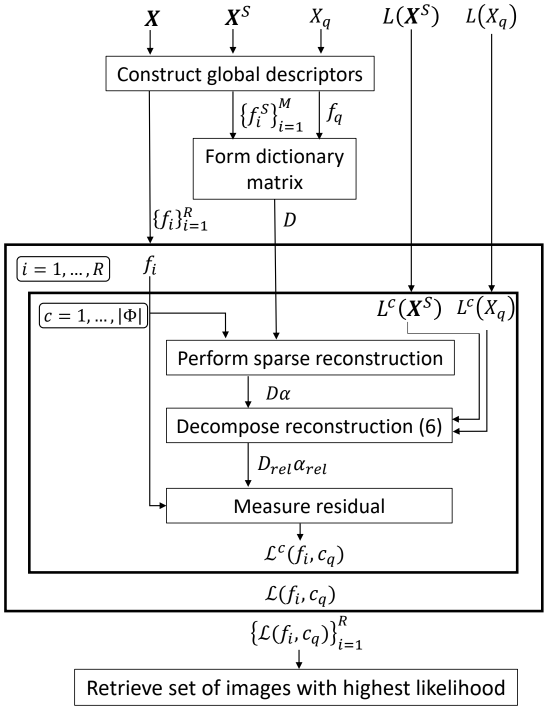


Fig. 4. Block diagram of the proposed multi-label retrieval method.

higher reconstruction performance, misleading the retrieval decision. Thus the choice for the images to be retrieved is based only on the residual related to c_q. In other words, rather than comparing the reconstruction performance with all categories, we compare the reconstruction performance of images using a single land-use category, namely that of the query image. The likelihood of an image $\boldsymbol{X_i}$ to belong to category c is therefore expressed as:

$$\mathcal{L}(\boldsymbol{f}_i, c) = 1 - \frac{\|\boldsymbol{f}_i - \mathbf{D}_c \boldsymbol{\alpha}_c\|}{\|\boldsymbol{f}_i\|} = 1 - \frac{r(\boldsymbol{f}_i, c)}{\|\boldsymbol{f}_i\|} \tag{5}$$

where the measure of likelihood $\mathcal{L}(\boldsymbol{f}_i, c)$ would ideally be expected to be close to 1 if $\boldsymbol{f}_i$ belongs to c and close to 0 otherwise. Thus for any given query image, the whole archive is evaluated in terms of its sparse reconstruction performance with respect to the category of the query image and hence the retrieved images are chosen as those that maximize the likelihood function $\mathcal{L}(\boldsymbol{f}, c_q)$.

B. Proposed Multi-label RS Image Retrieval Method

If the training images are annotated by multi-labels, the retrieval problem becomes more arduous. Nevertheless it is possible to make use of the approach developed for single-label image retrieval by decomposing the multi-label problem into multiple binary classification problems. Our proposed method, whose outline has been given in Fig. 4, makes use of training images in $\boldsymbol{X}_S$ equipped with their multi-label annotations $\{\Phi_i\}_{i=1}^M$ to retrieve images from $\boldsymbol{X}$ similar to X_q. In this way we aim to find images from $\boldsymbol{X}$ whose label sets most closely match the label set of the query image.

For a given label ϕ_c, the training set can be categorized into two groups, those which contain ϕ_c, the relevant set (denoted as rel), and those which do not, the irrelevant set (denoted as irr). Thus we get a binary partition of the training set for each label ϕ_c where each partition is indexed by the subscripts $\sigma \in \{rel, irr\}$. Let $\mathbf{D}_{rel}^c$ be the matrix made up of global descriptors of RS images from the relevant set ϕ_c, $\mathbf{D}_{irr}^c$ be the matrix formed by global descriptors from the irrelevant set and $\mathbf{D}$ be the matrix formed global descriptors from the whole training set. Given a label of interest ϕ_c, the sparse reconstruction of any image by $\mathbf{D}$ can be decomposed into the relevant and irrelevant components:

$$\boldsymbol{f}_i \approx \mathbf{D}\boldsymbol{\alpha} = \mathbf{D}_{rel}^c \boldsymbol{\alpha}_{rel}^c + \mathbf{D}_{irr}^c \boldsymbol{\alpha}_{irr}^c \tag{6}$$

This label-based partitioning of the training set can be used to estimate the likelihoods of an image to belong to the relevant and irrelevant subsets, where the likelihoods are computed as follows:

$$\mathcal{L}^c(\boldsymbol{f}_i, \sigma) = 1 - \frac{\|\boldsymbol{f}_i - \mathbf{D}_\sigma^c \boldsymbol{\alpha}_\sigma^c\|}{\|\boldsymbol{f}_i\|} = 1 - \frac{r_c(\boldsymbol{f}_i, \sigma)}{\|\boldsymbol{f}_i\|}$$

where $\sigma \in \{rel, irr\}$. Since the presence of a label is more decisive than the absence of it, we can limit our interest to the likelihood of rel and express label likelihood simply as:

$$\mathcal{L}^c(\boldsymbol{f_i}) = \mathcal{L}^c(\boldsymbol{f_i}, rel) = 1 - \frac{\|\boldsymbol{f_i} - \mathbf{D}_{rel}^c \boldsymbol{\alpha}_{rel}^c\|}{\|\boldsymbol{f_i}\|} \tag{7}$$

Using this function, the likelihood of both the query image X_q and the retrieved image X_r to contain a label ϕ_c would be expressed as $\boldsymbol{L}^c(X_q) \cdot \mathcal{L}^c(\boldsymbol{f}_r)$, where $\boldsymbol{L}(X_q)$ is the binary label vector as defined in Section II.A and $\boldsymbol{f}_r$ is the global descriptor of image X_r. Thus a measure of the shared query label likelihoods can be given as:

$$\sum_{c=1}^{K} \boldsymbol{L}^c(X_q) \cdot \mathcal{L}^c(\boldsymbol{f}_r) \quad \text{or} \quad \langle \boldsymbol{L}(X_q), \boldsymbol{\mathcal{L}}(\boldsymbol{f}_r) \rangle$$

Similarly, the likelihood of labels jointly absent would be given by the complementary vectors as $\langle \overline{\boldsymbol{L}}(X_q), \overline{\boldsymbol{\mathcal{L}}}(\boldsymbol{f}_r) \rangle$ where $\overline{\boldsymbol{L}}(X_q) = \mathbf{1} - \boldsymbol{L}(X_q)$ and $\overline{\boldsymbol{\mathcal{L}}}(\boldsymbol{f}_r) = \mathbf{1} - \boldsymbol{\mathcal{L}}(\boldsymbol{f}_r)$ with $\mathbf{1}$ the vector whose entries are all equal to 1. The overall maximization function can then be expressed as:

$$\boldsymbol{f}_r = \arg\max_{\boldsymbol{f} \in \boldsymbol{F}} \left(\langle \boldsymbol{L}(X_q), \boldsymbol{\mathcal{L}}(\boldsymbol{f}) \rangle + \gamma \langle \overline{\boldsymbol{L}}(X_q), \overline{\boldsymbol{\mathcal{L}}}(\boldsymbol{f}) \rangle \right) \tag{8}$$

where γ is a constant that determines the trade-off between retrieving all relevant labels and retrieving none of the irrelevant ones. Images retrieved using γ close to 0 are likely to contain most of the labels of the query image but also have many irrelevant ones while γ close to 1 would result in images that contain a smaller number of irrelevant labels but at the cost of a smaller number of relevant ones.

C. Weighting of Dictionary Words

We conjecture that the retrieval performance can be improved by making use of the sensitivity of the sparse reconstruction-based method to different dictionary words. Recall that the global descriptor of an RS image, as input to the retrieval system, is the histogram, that is, the set of occurrence frequencies, of the visual words. Some visual words are observed exclusively in images containing certain labels, and hence these words provide more significant information in determining the labels of those images. On the other hand, words whose occurrences are spread across many categories, offer little insight in determining the RS labels and their role in the decision process should be minimized. To this effect, we propose to scale words according to their significance based on a conditional entropy-based strategy. In the case of the category-based (i.e. single label) annotation of images, we achieve this as follows:

Let the conditional entropy of an RS image category given an occurrence of a word w_k be expressed as follows: $H(c|w_k)$. The conditional entropy is minimized if occurrences of w_k are limited to a single category, and thus this word becomes extremely significant. Conversely, maximum conditional entropy is achieved if word occurrences are spread evenly across all categories, thus the word has no value in determining image categories. We express the significance of a word w_k, also to be used as its weight, by the following formulation:

$$g(w_k) = log(C) - H(c|w_k). \tag{9}$$

Here $log(\text{C})$ corresponds to maximum value of entropy and therefore $g(w_k)$ is always non-negative. This corresponds to the amount of class uncertainty that word removes. The derivation for the computation of word entropy is given in Appendix A.

In the case of the multi-label RS image retrieval problems, where each label is treated separately, this approach can be implemented using binary labels, *relevant* versus *irrelevant*, defined in relation to the label of interest. Given any label of interest, the training set is partitioned into the two categories of relevant and irrelevant images (i.e. respectively images that contain the label and those that do not) and the algorithm presented for single-label retrieval is applied on partition. Therefore the multi-label application requires the computation of a separate set of word weights for each label, which is reasonable considering that a visual word could be strongly linked to a specific label, but might be completely unrelated to another one.

Given the diagonal matrix made up of the values of word weights $\mathbf{G} = diag\{g(w_1), \ldots, g(w_K)\}$, the sparse reconstruction formulation is modified (cf. (1)) by multiplying global descriptor entries by its corresponding weight:

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^M} \|\mathbf{G}\boldsymbol{f}_i - (\mathbf{G}\mathbf{D})\boldsymbol{\alpha}\| \text{ such that } \psi(\boldsymbol{\alpha}) \leq \eta \tag{10}$$

giving the label likelihood vector as:

$$\mathcal{L}^c(\boldsymbol{f_i}) = \mathcal{L}^c(\boldsymbol{f_i}, rel) = 1 - \frac{\|\mathbf{G}\boldsymbol{f_i} - (\mathbf{G}\mathbf{D}^c_{rel})\boldsymbol{\alpha}^c_{rel}\|}{\|\boldsymbol{f_i}\|}$$

V. Dataset Description and Experimental Setup

A. Dataset Description

The performance of the proposed RS CBIR system has been evaluated via experiments conducted on two benchmark archives. The first one is the widely used UC Merced Land Use benchmark archive consisting of 2100 images selected from aerial orthoimagery [31]. Each image is of size 256x256 pixels with high spatial resolution of 30 cm per pixel. Images are grouped into 21 different categories, where each image belongs to a single category. The category labels list as agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, and tennis court. Each category comprises 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. Examples of images and their single high-level category labels are given in Figure 5. More detailed information on this archive can be found in [31].

The second one is a benchmark archive consisting of 216 image tiles obtained by fragmenting a hyperspectral image acquired by EO-1 Hyperion sensor over surroundings of Ankara, Turkey [32], hence we refer to it as the Ankara archive. The images possess 119 spectral bands with a spatial resolution of 30 meters per pixel. Images are of size 63x63 pixels, and are annotated by low-level land cover class labels (i.e. multi-labels) from a set of 29 designated classes. However, since many of these labels are either present in nearly all images or occur in very few of them, we have only considered those labels for which there was sufficient diversity within the archive to enable a meaningful performance evaluation. Specifically, we have discarded all labels that are present in less than 15 images or in more than 200 images out of the total 216. Accordingly we have been left with 16 labels out of the original set of 29 labels. These specific labels are: arid soil, rocky terrain, tree, crop (type-A), crop (type-B), crop (type-D), red roofing, metal roofing, white roofing, blue roofing, membrane roofing, concrete roofing, unpaved road, asphalt pavement, grass (type-C), and pool. Example images are shown in Figure 6 with their multi-label information. For detailed information on this archive, we refer the Reader to [32].

B. Experimental Setup

For the UC Merced Land Use archive, each image is used as the query for only one image retrieval trial, resulting in a total of 2100 trials. A separate training set is randomly formed for each trial such that all the 21 categories are guaranteed to appear. These training sets contain 44 images, 4 of which belong to the category of the query image and the remaining 40 being formed of 2 randomly selected images per each of the 20 other categories. Note that the proposed system does not require to redefine the training set when the query image is varied. Random training sets are selected for each query only to show the robustness of our proposed system with respect to different training sets. For each such query image trial, we conducted experiments with different strategies in order to compare performances. We extract the local SIFT and BoSV descriptors on image windows of size 20x20. The SIFT descriptors are extracted from the first principal components, which are chosen to account for more than 90% of the total

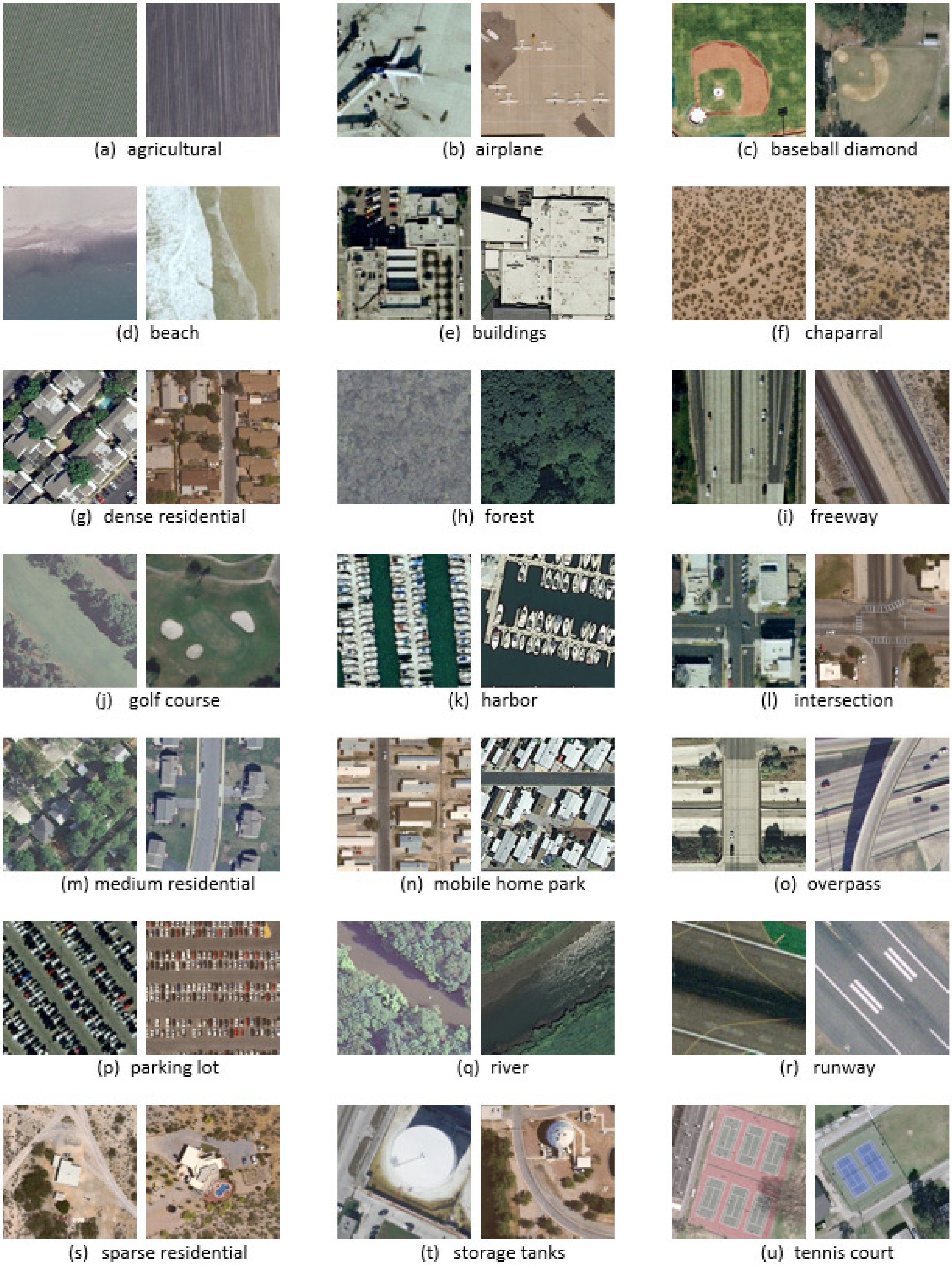
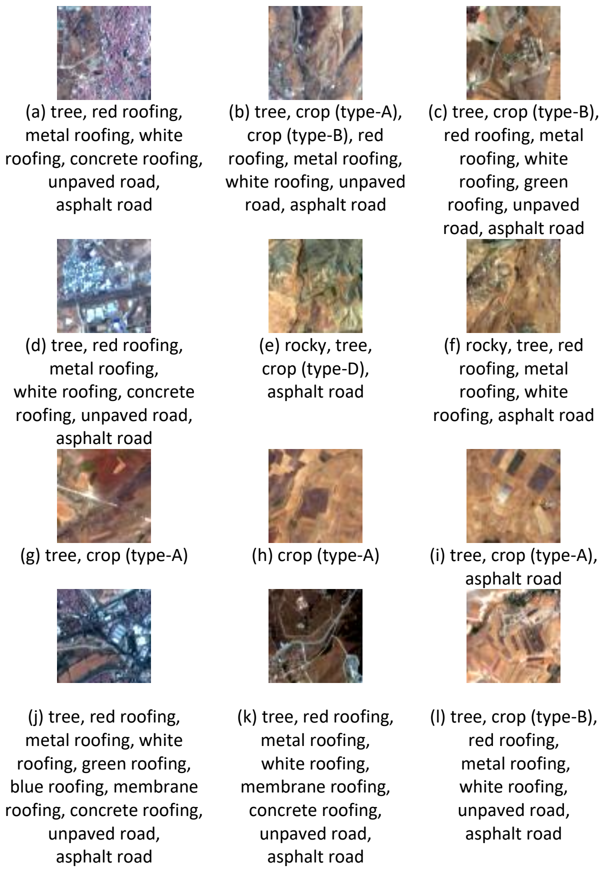


Fig. 5. Examples of images and their single high-level category labels in the UC Merced Land Use archive



(a) tree, red roofing, metal roofing, white roofing, concrete roofing, unpaved road, asphalt road
(b) tree, crop (type-A), crop (type-B), red roofing, metal roofing, white roofing, unpaved road, asphalt road
(c) tree, crop (type-B), red roofing, metal roofing, white roofing, green roofing, unpaved road, asphalt road
(d) tree, red roofing, metal roofing, white roofing, concrete roofing, unpaved road, asphalt road
(e) rocky, tree, crop (type-D), asphalt road
(f) rocky, tree, red roofing, metal roofing, white roofing, asphalt road
(g) tree, crop (type-A)
(h) crop (type-A)
(i) tree, crop (type-A), asphalt road
(j) tree, red roofing, metal roofing, white roofing, green roofing, blue roofing, membrane roofing, concrete roofing, unpaved road, asphalt road
(k) tree, red roofing, metal roofing, white roofing, membrane roofing, concrete roofing, unpaved road, asphalt road
(l) tree, crop (type-B), red roofing, metal roofing, white roofing, unpaved road, asphalt road

Fig. 6. Example images and their low-level land cover class labels in the Ankara archive

variance (see Section III.A). The size of the codebook of spectral values used for the BoSV was chosen as 100. We considered a codebook of size 500 for the SIFT descriptors, 100 for raw pixel descriptors (RPD) and 500 for the BoSV descriptors (i.e. SBoSV and EBoSV). Thus the resulting global descriptors are of length 100 for the raw pixel description, 500 for the SIFT or BoSV descriptions, 600 for the SIFT-RPD hybrid description, and 1000 for the SIFT-BoSV hybrid descriptions. The performance of the system for different sizes for the codebooks of spatial/spectral descriptors has also been evaluated. The codebook sizes that we have considered are 100, 250, 750 and 1000 in addition 500. For the hybrid description, weight ratio of descriptors is 0.8/0.2 in favor of spatial description, i.e. $\lambda_{spect} = 0.2$ in (4). This value has been chosen as a simple ratio (4 to 1) that achieves good performance. This has been selected after performing cross validation with different values of λ_{spect} ranging from 0.1 to 0.9 with increments of 0.1. The performance of the single label retrieval system is expressed in terms of the precision value for 15 retrieved images, where precision is simply the ratio of correctly relevant images (i.e. those belonging to the category of the query image) among all retrieved images.

For the Ankara archive, due to the lack of a satisfactory number of potential query images, each image is used as the query in 5 different trials of image retrieval, resulting in a total of 1080 trials. Similar to our testing approach for UC Merced Land Use archive, we have performed different trials on the Ankara archive with different and arbitrarily chosen training sets. Before each trial, a set of training images is randomly selected so as to satisfy a criterion of diversity for the training phase to be successful. Our criterion is such that for any label, there must be at least 5 images that contain that label and at least 5 others that do not. The size of the resulting training sets vary, but contain on the average 29 images. Given a query image and its training set, each experimental trial is conducted several times using different image descriptions and retrieval approaches for comparative evaluation. Considering the low spatial resolution, the local SIFT and BoSV descriptors are extracted on image windows of size only 8x8. However owing to the high spectral resolution, spatial descriptors are extracted from the first 3 principal components, which together contain more than 90% of total variance (see Section III.A). The size of the codebook of spectral values used for the BoSV was chosen to be 500. We considered a codebook of size 500 for the SIFT descriptors, the RPD and the BoSV descriptors. Thus the resulting global descriptors are of length 500 with purely spatial or spectral descriptions and 1000 with the hybrid descriptions. The spatio-spectral weight of hybrid description is 0.5/0.5, i.e. $\lambda_{spect} = 0.5$ in (4), selected as a simple ratio with good performance. Once again, we have reiterated our experiments for various sizes of the descriptor codebooks, ranging from 100 to 1000, as well as for different values of the spectral weight parameter λ_{spect} in the gamut from 0.1 to 0.9 with 0.1 increments.

The performance of the proposed multi-label retrieval system is expressed in terms of 4 metrics for 15 retrieved images. Given that Φ is the set of all labels, Φ_q is the set of labels associated with the query image and Φ_r is the set of labels associated with a retrieved image X_r, these metrics are expressed as follows:

Precision:

$$\frac{1}{|\boldsymbol{X}_r|} \sum_{r:X_r \in \boldsymbol{X}_r} \frac{|\Phi_q \cap \Phi_r|}{|\Phi_r|}$$

Accuracy:

$$\frac{1}{|\boldsymbol{X}_r|} \sum_{r:X_r \in \boldsymbol{X}_r} \frac{|\Phi_q \cap \Phi_r|}{|\Phi_q \cup \Phi_r|}$$

Recall:

$$\frac{1}{|\boldsymbol{X}_r|} \sum_{r:X_r \in \boldsymbol{X}_r} \frac{|\Phi_q \cap \Phi_r|}{|\Phi_q|}$$

Hamming Loss:

$$\frac{1}{|\boldsymbol{X}_r|} \sum_{r:X_r \in \boldsymbol{X}_r} \frac{|\Phi_q \setminus \Phi_r| + |\Phi_r \setminus \Phi_q|}{|\Phi|}$$

where $\boldsymbol{X}_r$ corresponds to the set of retrieved images. Note that while we report Precision, Recall, Accuracy and Hamming Distance in Table IV for the multi-label images, for the single-label case the definitions of Precision, Recall and Accuracy for multi-label retrieval all reduce to the same metric. Hence in Table 1 we only report Precision results for the single-label retrieval trials.

To compare our single label and multi-label image retrieval methods with the SVM-based retrieval method, SVM classifiers have been trained to evaluate label likelihood based on the prediction scores of SVM classification. The classifiers have been trained using the radial basis function kernel with the regularization parameter obtained by 5-fold cross validation. For the single-label case, training has been done using a one-versus-all approach, with images from the query category on one side and all other training images on the other. For multi-label image retrieval, SVMs were trained using an approach comparable to our multi-label sparsity-based retrieval method (see IV.B). For each label of interest the training set is partitioned into 2 groups: relevant images (those that contain the label) and irrelevant images (those that do not). A separate SVM classifier was trained according to these partitions for each label, resulting in a parallel architecture of the 16 SVM classifiers used to assign the label likelihood vector.

Computational times required for image description have been computed by performing the same process on a set of 100 randomly selected images for each choice of description and taking the average value. A similar approach has been used for computational times required for image retrieval, where the average duration of 100 trials with randomly selected queries has been measured. All experiments are implemented via MATLAB on a PC with Intel Xeon E5 1.8GHz processor and 36GB RAM.

VI. Experimental Results

We performed experiments aimed to analyze: i) the effectiveness of spatial, spectral and spatio-spectral (i.e., hybrid) image descriptions, ii) the success of the proposed spectral descriptors, iii) the effectiveness of the proposed sparse reconstruction-based retrieval system, and finally iv) the performance of the proposed word weighting method evaluated in the framework of single-label and multi-label image retrieval problems.

A. Results of Single-label Image Retrieval on the UC Merced Land Use Archive

In this subsection, we assess the effectiveness of the proposed system in the retrieval of images from the UC Merced Land Use archive. In the first set of experiments, we evaluate the performance of our proposed spatial and spectral descriptors individually and then jointly. To this end, Table I shows the average precision obtained when images are characterized: 1) only by the SIFT-based spatial description; 2) only by the proposed spectral description based on RPD, SBoSV and EBoSV, respectively (see Fig. 2), and 3) hybrid descriptors obtained by combining the SIFT-based descriptor with the RPD, SBoV and EBoSV. Computational time required for each description method is also given in this Table. The performance of each hybrid description method has been given below that of spectral description based on the same spectral descriptor to stress the contribution of spatial descriptors to each of the spectral descriptors. All computational time measures have been given relative to the duration $T = 3.03$ seconds/image of pure spatial description. Note that these results have been obtained using our sparse reconstruction-based retrieval algorithm.

TABLE I
PERFORMANCE COMPARISON OF PROPOSED IMAGE DESCRIPTIONS: UC MERCED LAND USE ARCHIVE

Proposed Descriptors	Precision (%)	Time
SIFT	77.8	T
RPD	63.5	**0.16T**
SIFT+RPD	**85.0**	1.02T
SBoSV	71.6	0.81T
SIFT+SBoSV	84.8	1.66T
EBoSV	70.0	3.87T
SIFT+EBoSV	84.9	4.54T

From this Table one can see that the SIFT-based spatial description considerably outperforms spectral description with a difference of at least 6%. This is expected since these archive images have high spatial resolution, but very low spectral resolution. The fusion of information from both types of descriptors, however, results in a performance higher than both spectral and spatial descriptions. As an example, the fusion of the SIFT with RPD achieves 85% precision compared to the 77.8% precision obtained by using only the SIFT-based description. This shows that the spatial description does not capture all the available information and there is crucial information resting associated with the spectral components of the RGB images. Furthermore it is remarkable that this increase in performance by the addition of RPD-based spectral description is achieved with a negligible increase in computation time, as the SIFT+RPD hybrid takes only 2% longer than the purely spatial SIFT-based description. Comparing spectral descriptors among themselves we observe that each one achieves similar performance when used in conjunction with spatial features. However the BoSV descriptors (i.e., SBoSV and EBoSV) are more robust against the lack of spatial description, achieving a performance difference of 7% against the RPD when used without spatial description. This is easily explained by the fact that the BoSV descriptors make use of contextual information, taking into account the neighborhood relations between pixels. Since contextual spectral information might not always be captured by the SIFT-based description, this does not turn out to be the case for this archive, where spectral information is limited to only three bands. Furthermore the difference in computational time between the RPD and the BoSV descriptors is significant, making the latter very inefficient for this case.

TABLE II
CONTRIBUTION OF THE PROPOSED WORD-WEIGHTING STRATEGY ON RETRIEVAL PERFORMANCE FOR THE UC MERCED LAND USE ARCHIVE

Results	SRR-noWW	SRR-WW
Precision(%)	83.4	**85.1**
Time	T	1.004T

In the second set of experiments, we evaluated the contribution of word weighting on the proposed sparse reconstruction-based retrieval (SRR) performance. Table II shows the precision and the average computational time obtained by including the proposed word weighting strategy in SRR (denoted as SRR-WW) and by excluding it (denoted as SRR-noWW).

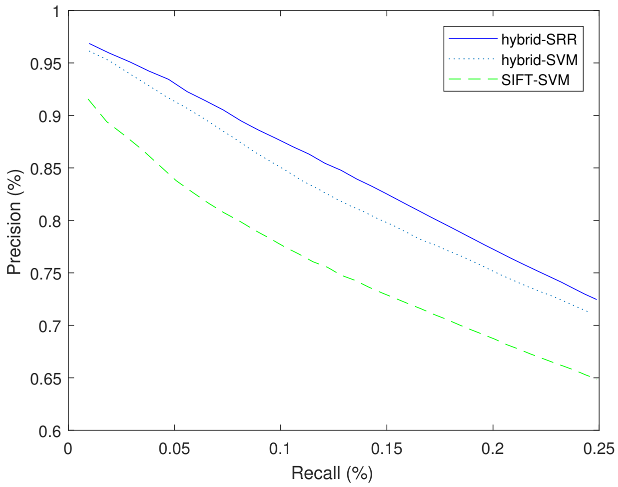


Fig. 7. Precision recall curve for SIFT-SVM, hybrid-SVM and the proposed hybrid-SRR

Computational time has been given relative to the duration $T = 2.013$ seconds/query of SRR-noWW. These results in Table II are obtained using the SIFT+SBoSV image descriptions. From this Table one can see that word weighting improves the performance of the retrieval system modestly, by 1.5% while costing a truly negligible increase in computation time of only 0.4%.

In the third set of experiments, we compared the performance of the standard SIFT description [23] (denoted as SIFT-SVM) with the method where both spatial and spectral content are taken into account (denoted as hybrid-SVM) as well as with the proposed retrieval method, where spatio-spectral description is used in conjunction with the sparse reconstruction-based retrieval method (denoted proposed hybrid-SRR). Note that the hybrid performances are based on expressions obtained using the SIFT+SBoSV image descriptions, the average precision obtained by each method is given in Table III. The precision-recall curves are given in Figure 7.

TABLE III
RESULTS OBTAINED BY THE SIFT-SVM, THE HYBRID-SVM, AND THE PROPOSED HYBRID-SRR ON THE UC MERCED LAND USE ARCHIVE

Results	SIFT-SVM	hybrid-SVM	proposed hybrid-SRR
Precision(%)	76.1	82.2	**85.1**

Experimental results demonstrate a considerable increase in performance using the SRR over the SVM-based retrieval method, with a difference of 3% of the hybrid-SRR system over the hybrid-SVM. Comparing the performance of the complete proposed system with novel spectral description against SVM retrieval with the SIFT-based description, the difference becomes even more striking, with the hybrid-SRR outperforming the SIFT-SVM by 9%. Computationally our implementation of the SRR on MATLAB takes significantly longer than the built-in tools available for SVM. The difference in computational time is almost one order of magnitude, however this problem can be overcome by considering a cluster-based parallel computing approach.

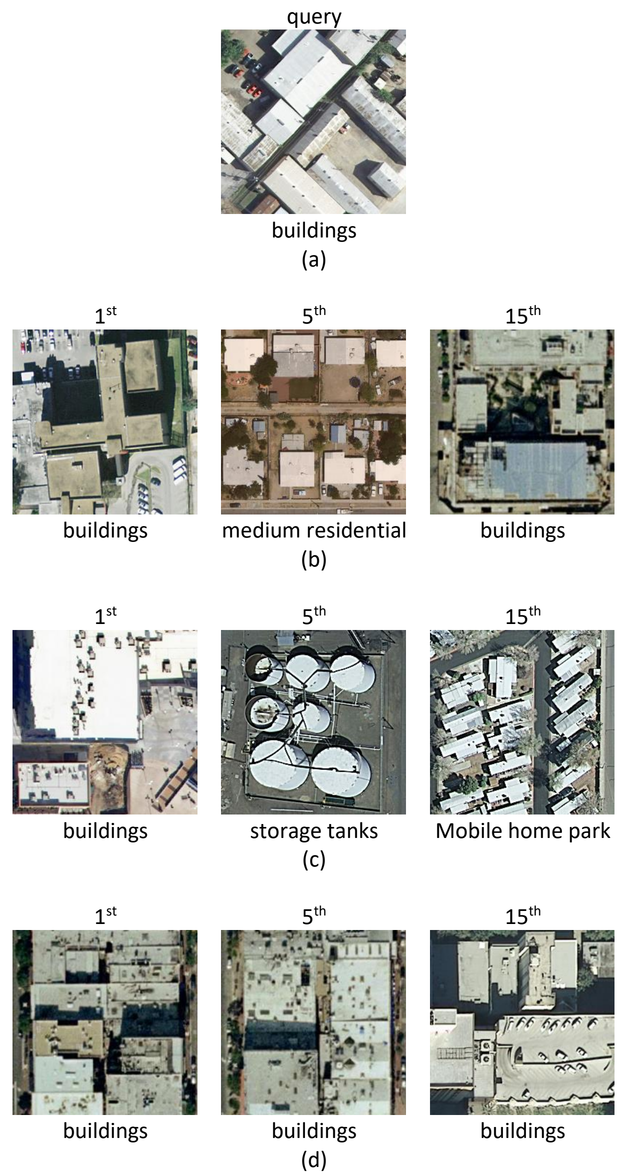


Fig. 8. Retrieval results from the UC Merced Land Use archive: (a) query image, (b) images retrieved by SIFT-SVM, (c) images retrieved by hybrid-SVM and (d) images retrieved by the proposed hybrid-SRR (corresponding category labels are reported below each image)

Fig. 8 shows an example of images retrieved by the SIFT-SVM, the hybrid-SVM and the proposed hybrid-SRR related to a query image (Fig. 8 (a)) taken from the buildings category. The retrieval order of each image is given above the related image and the category with which the image is associated is given below the related image. By visual analysis of several such results, we have concluded that the proposed system models more accurately the images associated with each query image and retrieves the more visually similar images from the archive. These observations suggest the following: The proposed hybrid-SRR retrieves images of with more similar content (all three from the buildings category), while the hybrid-SVM introduces more variation which in this case results in some erroneous results (such as an image of storage tanks being retrieved instead of an image of the desired category). For the SIFT-SVM on the other hand, we observe a significant variation in color, which is expected considering the lack of any spectral description scheme. Note the 5th retrieved image standing out with the brown color of soil dominating the image.

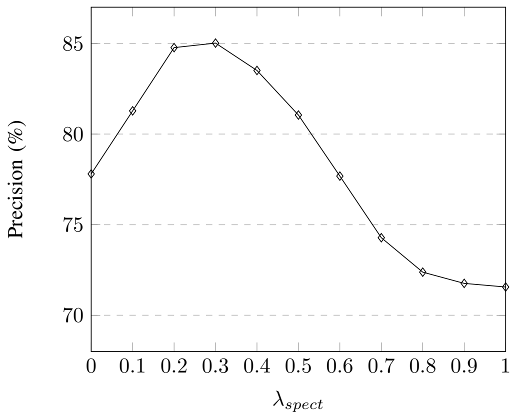


Fig. 9. Results obtained under different values of λ_{spect} on the UC Merced Land Use archive

In the fourth set of experiments, we explored the sensitivity of our system to the choice of the spectral weight coefficient λ_{spect}. We performed retrieval on the archive using values of λ_{spect} ranging from 0.1 to 0.9. The performance obtained under different values of λ_{spect} as well as by the purely spatial and purely spectral description methods (corresponding to $\lambda_{spect} = 0$ and $\lambda_{spect} = 1$ respectively) are given in Fig. 9. It can be seen from the Figure that the performance reaches a peak at $\lambda_{spect} = 0.2 \sim 0.4$. The lower performance of the system at high values of the spectral weight shows that spectral description itself is limited in expressive capability. Nevertheless its contribution to the system for a judiciously selected weight is significant. Note that the performance consistently increases with increasing weight before reaching its peak around $\lambda_{spect} = 0.25$. This shows that the choice of the spectral weighting parameter is indeed significant and the relative importance of spectral and spatial features must be assessed carefully. In our experiments this has been done empirically through cross validation.

In the fifth set of experiments, we compared the performance of our system by jointly varying the size of our spatial and spectral codebooks. The result of our experiments, using codebooks of size 100, 250, 500, 750 and 1000 are given in Fig. 10. We observe that, as expected, the system suffers in performance when the codebook is not large enough to accomodate the diversity in descriptors extracted from the archive. However it is also worth noting that beyond a certain point we do not observe any consistent gain by increasing the codebook size. Overall this suggests that the system performs well for a wide range of values of codebook size, as long as the codebook is larger than some critical threshold.

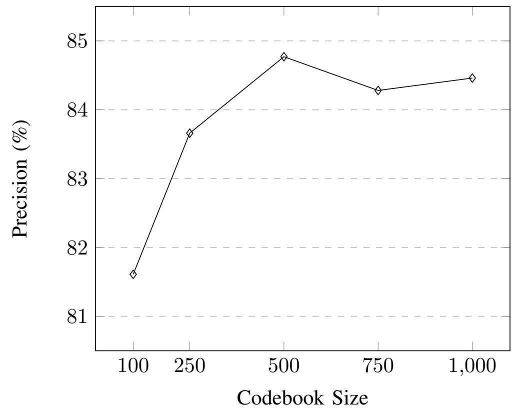


Fig. 10. Results obtained for different codebook sizes on the UC Merced Land Use archive

B. Results of Multi-label Image Retrieval on the Ankara Archive

In this subsection we evaluate the effectiveness of the proposed system in the retrieval of multi-label hyperspectral images from the Ankara archive. Similar to the previous subsection, the first set of experiments focuses on analyzing the performance obtained by our spatial and spectral descriptors. In Table IV we give the retrieval performance obtained by multi-band spatial descriptors (mSIFT), spectral descriptors (RPD, SBoSV and EBoSV) and hybrid description based on the joining of both types of descriptors (mSIFT-RPD, mSIFT-SBoSV and mSIFT-EBoSV). As in Table I, computational time measures are given relative to the duration $T = 2.17$ seconds/image of the pure spatial description. The reported results have been obtained using the SRR algorithm. The performance with the spectral description is, as expected, higher than that with spatial description, due to the poor spatial resolution. As an example EBoSV yields 3% higher precision than mSIFT. In this case, the fusion of information obtained by spatial description with that from spectral description does not bring any significant performance improvement. The precision of the spectral/hybrid pairs (RPD and RPD-mSIFT, SBoSV and mSIFT+SBoSV, EBoSV and mSIFT+EBoSV) are all very close with each hybrid description method obtaining at most 0.6% higher precision than the corresponding purely spectral method. This suggests that the very high spectral resolution does not leave room for information gain from spatial information, especially at such low resolutions. The same effect can be seen in the inefficiency of the BoSV (i.e. SBoSV and EBoSV) where contextual information provides little benefit. It would be informative to find out the spatial resolution level beyond which one starts to observe a discernible improvement in the use of the BoSV descriptors. As expected, we note that accuracy, recall and hamming loss consistently mirror the same trends in performance displayed by the precision metric.

Our second set of experiments was performed to investigate the contribution of word weighting to our retrieval system. The results for SRR-noWW and SRR-WW are given in Table V in addition to computational times, given relative to the duration $T = 2.258$ seconds/query of SRR-noWW. From

TABLE IV
PERFORMANCE COMPARISON OF DIFFERENT PROPOSED DESCRIPTIONS: ANKARA ARCHIVE

Image Descriptors	Precision(%)	Accuracy(%)	Recall(%)	Hamming Loss(%)	Time
mSIFT	67.9	55.2	71.9	20.5	T
RPD	72.3	61.0	78.4	18.0	**0.28T**
mSIFT-RPD	**72.7**	61.5	78.9	**17.7**	1.23T
SBoSV	70.5	60.7	79.8	18.8	1.33T
mSIFT-SBoSV	70.9	61.3	80.2	18.5	2.27T
EBoSV	70.9	61.1	79.9	18.6	6.00T
mSIFT-EBoSV	71.5	**61.7**	**80.5**	18.0	6.68T

this Table one can see that word weighing weighting further improves the performance of the retrieval system, resulting in a difference in performance ranging from 1.7% in precision to 4.5% in accuracy, with computational time remaining almost unchanged.

TABLE V
CONTRIBUTION OF THE PROPOSED WORD-WEIGHTING ON RETRIEVAL PERFORMANCE FOR THE ANKARA ARCHIVE

Results	SRR-noWW	SRR-WW
Precision(%)	69.2	**70.9**
Accuracy(%)	56.8	**61.3**
Recall(%)	74.8	**80.2**
Hamming Loss(%)	20.8	**18.5**
Time	T	1.01T

In our third set of experiments we evaluated the performance of the proposed SRR-based system versus the SVM-based methods. To this end Table VI shows the performance obtained using the mSIFT-SVM, the hybrid-SVM and the proposed hybrid-SRR. The reported results are obtained using the mSIFT+SBoSV for image description. Experimental results demonstrate the proposed hybrid-SRR achieving a considerable increase in performance over the hybrid-SVM approach with a 2.7% difference in precision in favor of SRR. Note that, owing to the low spatial resolution of the archive, the necessity of spectral description is even more striking in this case, with the mSIFT-SVM obtaining only 61% precision as compared to the 71% of the hybrid-SRR. By analysing Table one can also observe that similar behavior is obtained under different performance metrics. As one example we can cite the 8% difference in accuracy of hybrid-SRR over hybrid-SVM and the more significant 16% difference in accuracy it achieved over mSIFT-SVM. Finally note that the difference in computational time between our implementation of the SRR and the available built-in tools for SVM persists, where the hybrid-SRR takes an order of magnitude longer than the SIFT-SVM or the hybrid-SVM.

TABLE VI
RESULTS OBTAINED BY THE MSIFT-SVM, THE HYBRID-SVM AND THE PROPOSED HYBRID-SRR ON THE ANKARA ARCHIVE

Results	mSIFT-SVM	hybrid-SVM	proposed hybrid-SRR
Precision(%)	60.7	68.2	**70.9**
Accuracy(%)	45.2	53.0	**61.3**
Recall(%)	61.4	69.4	**80.2**
Hamming Loss(%)	26.8	22.5	**18.5**

Fig. 11. Retrieval results from the Ankara archive: (a) query image, (b) images retrieved by the mSIFT-SVM, (c) images retrieved by the hybrid-SVM and (d) images retrieved by the proposed hybrid-SRR (corresponding multi-labels are reported below each image). Retrieved image labels that are not associated with the query image are given in italics.

In Figure 11 we present a retrieval trial with the query image (Fig. 11(a)) and the sets of 3 retrieved images obtained by the mSIFT-SVM (Fig. 11(b)), the hybrid-SVM (Fig. 11(c)) and the hybrid-SRR (Fig. 11(c)), all given as true color composite images. The given results visually demonstrate the proposed hybrid-SVM system's capability in retrieving images very close to the query image. In Fig. 11 (d) the 5th image retrieved by the hybrid SRR contains a label set nearly identical to that of the query image, containing all relevant labels and with a single irrelevant label:'unpaved road'. This is remarkable considering the high number of labels associated with both images. The 1st and 15th images retrieved by the proposed system are not as accurate (missing multiple labels of the query image). Nevertheless it should be noted that the 1st image captures exactly the agricultural composition, which is dominant characteristic of the query, while the 15th image presents an image containing labels of buildings and farmlands, which corresponds to the composition of the query. We note that none of the images retrieved by the other methods include any labels associated with crops. We observe images retrieved by the hybrid-SVM system to show significant deviation from the query image in image content. In Fig. 11 (c) all images contain multiple unrelated labels of different types of roofing. Furthermore we observe the lack of any type of label associated with agriculture, suggesting images corresponding to areas of settlements rather than farmlands. This is confirmed by visual inspection of the images. Results retrieved by the mSIFT-SVM in Fig. 11 (b) demonstrate a more significant shortcoming of the method. Inspection of the true color composite images shows the images by the SIFT-SVM being visually very different from the query image, suggesting them to have hyperspectral content quite distinct from that of the query image. Although observing the true color composite image results is not by itself enough to make deductions on the systems capacities in distinguishing hyperspectral characteristics, this observation is in line with the results observed on the previous UC Merced Land Use archive and with what we would expect. Nevertheless the retrieved images are similar to those retrieved by the hybrid-SVM in image content, as both contain labels that suggesting denser settlements and lack of farmland. We observe that labels associated with different types of buildings, such as membrane roofing and white roofing, are consistently present in the retrieved images.

In the fourth set of experiments we explored the sensitivity of our system to the choice of the spectral weight coefficient λ_{spect}. We performed retrieval on the archive using values of λ_{spect} ranging from 0.1 to 0.9. The performance obtained under different values of λ_{spect} as well as by the purely spatial and purely spectral description method (corresponding to $\lambda_{spect} = 0$ and $\lambda_{spect} = 1$ respectively) have been given in Fig. 12. The Figure shows all performance metrics following the same trend, implying a common conclusion. The variation in performance as a function of λ_{spect} clearly shows the significance of spectral information for the case of hyperspectral images, as the performance starts with low spectral weight but rapidly increases and eventually reaches a saturation point as spectral weight is increased. While the performance metrics do not change much once λ_{spect} is at least $0.4 \sim 0.5$, it must be noted that there is a slight decline around the point $\lambda = 0.5$. This slight loss in performance reflects the contribution of spatial description to the overall description pipeline, even in the case of the hyperspectral archive with low spatial resolution.

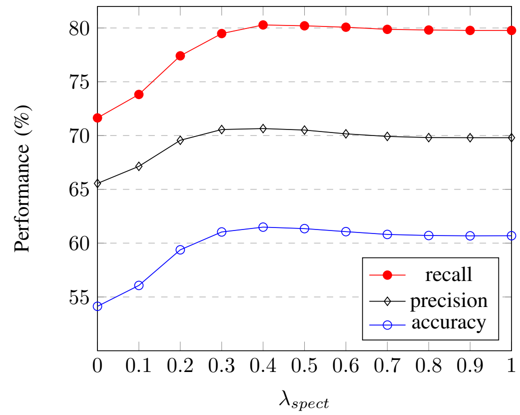


Fig. 12. Results obtained under different values of λ_{spect} on the Ankara archive.

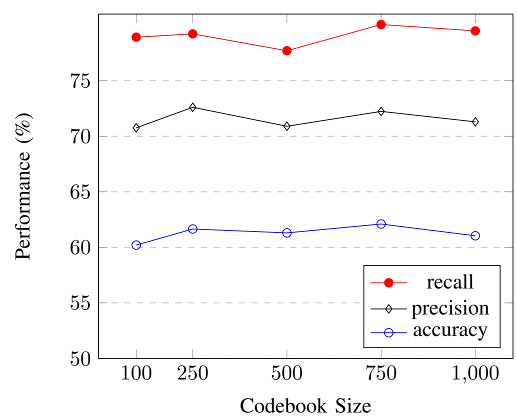


Fig. 13. Results obtained for different codebook sizes on the Ankara archive.

In the fifth set of experiments we compare the performance of our system by jointly varying the size of our spatial and spectral codebooks. The result of our experiments, using codebooks of size 100, 250, 500, 750 and 1000 are given in Fig. 13. Once again, the Figure shows a slight increase in performance as we increase the codebook size. However we do not observe a consistently upwards trend but instead the performance fluctuates as the codebook size is increased. This behavior matches the one observed in Fig. 10 beyond the point of critical size. Hence it suggests that the critical minimum size for codebooks is smaller for the Ankara archive, presumably not much greater than 100. This is explained by the underlying uniformity of the archive: the Ankara archive consists of images all belonging to a single region while the UC Merced Land Use archive is made up of images taken from numerous areas with a significantly greater diversity in geographical features. Hence it is expected that our method performs reasonably well on this archive even with relatively small codebook sizes.

VII. Conclusion

In this paper a novel content-based image retrieval system has been presented for remote sensing archives. The proposed system models and exploits conjointly the spectral and the spatial information content of images, using two image description pipelines, both based on separate BoVW approaches. To this end three varieties of local spectral descriptors have been developed with increasing levels of context information: 1) RPD, which simply expresses the frequency of discretized pixel values in the image; 2) SBoSV, which not only captures how frequent pixel values are, but also their co-occurence statistics; and finally 3) EBoSV, which also takes into account the spatial distribution of pixel values as well as their joint frequencies of occurrence. For spatial description we have adopted a multi-band implementation of the SIFT approach. The outcomes of the description pipelines are then used for image retrieval through a sparse reconstruction-based retrieval method developed as an adaptation for single and multi-label image retrieval of an existing classification method. Finally, a measure of visual word significance is introduced to weight the BoVW entries to enhance retrieval performance.

In order to evaluate the proposed system, we considered two benchmark archives. Experimental results verify the necessity of spectral description concomitantly with spatial description for RS image retrieval. Judicious use of spectral description greatly improves the retrieval performance whenever multi-band information is available. Spectral description proves to be especially valuable at low spatial resolutions. While all the three proposed local spectral descriptors have comparable contributions to the retrieval performance, we believe the most suitable one can be chosen by matching it with the spatial-spectral properties of the archive. Experiments also demonstrate that the proposed sparse reconstruction-based retrieval method, called SRR, outperforms SVM in all cases, and that attributing weights to the BoVWs (actually to their histogram bins) improves performance. The motivations behind the proposed system are: 1) to achieve more accurate description of the information content of images enabled by the conjunction of spatial and spectral features, 2) to develop a system capable of single-label and multi-label retrieval that incorporates the strengths of sparse reconstruction-based methods, and 3) to enhance image expressions by highlighting their more informative features.

For our final remarks, we would like to point out that we have implicitly assumed that training sets are representative of all the underlying categories/labels as observed in the images. In other words, our system does not allow for discovery of undefined categories. In case of a partially labeled archive, it is critical to be able to exploit all the available data, labeled or unlabeled. As a future development, we plan to integrate an active learning scheme within our system in order to make use of unlabeled samples. It is worth noting that for large-scale operational RS CBIR problems, the proposed system may require high retrieval time as any other retrieval system. This can be solved by considering a cluster-based parallel computing approach for the implementation of the proposed system. For further analysis and development, we plan to exploit nonlinear descriptors and explore alternative ways to integrate spatial and spectral characteristics. For example, one such alternative would be the early fusion of local spatial and spectral information to obtain local hybrid descriptors. Finally a comparative performance study of spatial descriptors could be of interest in itself. A possible step in this direction can be the use features extracted via deep networks [35].

Appendix A
Conditional word entropy

The presented method makes use of the conditional entropy $H(c|w_k)$ as a measure of the significance of the occurrence of any word w_k in inferring the category c of an image. This requires an expression for $P(c|w_k)$, the conditional probability of an image to belong to category c given the occurrence of visual word w_k, which can be obtained using the Bayes' formula:

$$P(c|w_k) = \frac{P(w_k|c)P(c)}{P(w_k)}$$

$P(w_k|c)$, the conditional probability of a word occurrence in an image to be that of w_k given the category c of the image, can be expressed as the average frequency of occurrence of any word w_k in images belonging to category c, which we note with the shorthanded notation $mean(f_i^k|c)$:

$$P(w_k|c) = \frac{1}{|\boldsymbol{X}_c|} \sum_{X_i \in \boldsymbol{X}_c} f_i^k = mean(f_i^k|c)$$

Assuming a lack of information regarding the prior probability, we suppose categories to be equally probably:

$$P(c) = \frac{1}{C}$$

Then the marginal probabilities of word occurrences can be computed as:

$$P(w_k) = \sum_{c=1}^{C} P(w_k|c)P(c) = \frac{1}{C} \sum_{c=1}^{C} mean(f_i^k|c)$$

By Bayes' theorem this gives:

$$P(c|w_k) = \frac{P(w_k|c)P(c)}{P(w_k)} = \frac{mean(f_i^k|c)}{\sum_{c'} mean(f_i^k|c')}$$

which is then used to compute the conditional entropy:

$$H(c|w_k) = -\sum_{c=1}^{C} P(c|w_k) \log H(c|w_k)$$

Appendix B
Notation and symbols

An list of the notation and symbols used throughout the paper is given in Table VII.

Acknowledgment

This work was supported by the European Research Council under the ERC Starting Grant BigEarth-759764.

TABLE VII
TABLE OF SYMBOLS

Symbol	Concept
$X_i \in \boldsymbol{X}$	Images
$X_i^S \in \boldsymbol{X}_S$	Annotated images (i.e. training set)
$X_{c_i}^S \in \boldsymbol{X}_c$	Annotated images exclusively belonging to class c ($\boldsymbol{X}_c \subset \boldsymbol{X_S}$)
$\boldsymbol{f}_i = \{f_i^1, \ldots, f_i^K\} \in \boldsymbol{F}$	Global descriptors i.e. histogram vectors where f_i^k corresponds to the frequency of occurence of w_k in X_i
$(\boldsymbol{f}_i)^S \in \boldsymbol{F}_S$ and $(\boldsymbol{f}_{c_i})^S \in \boldsymbol{F}_c$	Same as in $X_i^S \in \boldsymbol{X}_S$ and $X_{c_i}^S \in \boldsymbol{X}_c$
$c \in \{1, \ldots, C\}$	Class index
$R = \|\boldsymbol{X}\|$	Archive size
$M = \|\boldsymbol{X}_S\|$	Training set size
$M_c = \|\boldsymbol{X}_c\|$	Number of training images belonging to class c
$\phi_c \in \Phi$	Label of class c in the set of all labels
$\Phi_i \subset \Phi$	Set of labels associated with image X_i
$\boldsymbol{L}(X_i^S) \in \{0,1\}^{\|\Phi\|}$	Binary label vector for annotated image X_i^S
$\boldsymbol{\mathcal{L}}(\boldsymbol{f}_i) \in [0,1]^{\|\Phi\|}$	Label likelihood vector for non-annotated image $\boldsymbol{X}_i$
$w_k \in \boldsymbol{W}$	Visual words
N	Number of spectral bands
n	Number of bands after dimension reduction
$\widetilde{X}_i$	Image after dimension reduction
$\mu_{(i,j)}^b \in [0,1]^{128}$	SIFT descriptor obtained at j-th location b-th band of image $\widetilde{X}_i$
$\mu_{(i,j)} \in [0,1]^{128 \times n}$	multi-band SIFT descriptor obtained at j-th location of $\widetilde{X}_i$
$\xi_{(i,j)}$	Spectral descriptor obtained at j-th location of X_i
$\{\widehat{\boldsymbol{\mu}}_1, \ldots, \widehat{\boldsymbol{\mu}}_{K_{spat}}\}$	Set of spatial visual words
$\{\widehat{\boldsymbol{\xi}}_1, \ldots, \widehat{\boldsymbol{\xi}}_{K_{spect}}\}$	Set of spectral visual words
$\mathbf{D}_c = [\boldsymbol{f}_{c_1}, \ldots, \boldsymbol{f}_{c_{M_c}}]$	Dictionary made up of global descriptors associated with class c
$\mathbf{D}_{rel}^c$ (and $\mathbf{D}_{irr}^c$)	Dictionaries made up of global descriptors associated (and not associated) with label ϕ_c
$\mathbf{D}$	Dictionary made up of all global descriptors from training set
$\boldsymbol{\alpha}$, $\boldsymbol{\alpha}_c$, $\boldsymbol{\alpha}_{rel}^c$	Sparse reconstruction coefficient vectors of corresponding dictionaries
γ	constant determining retrieval trade-off of non-annotated/undesired labels
$\mathbf{G} = diag(g(w_1), \ldots, g(w_K))$	Diagonal matrix containing word weights

REFERENCES

[1] O. E. Dai, B. Demir, B. Sankur, L. Bruzzone, "A Novel System for Content Based Retrieval of Multi-Label Remote Sensing Images," in *Int. Geoscience and Remote Sensing Symp.*, Texas, USA, 2017.

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

[3] T. Bretschneider, R. Cavet, and O. Kao, "Retrieval of remotely sensed imagery using spectral information content," in *Proc. IEEE Int. Geoscience Remote Sensing Symp.*, Toronto, Canada, 2002, pp. 2253—2255.

[4] T. Bretschneider and O. Kao, "A retrieval system for remotely sensed imagery," in *Proc. Int. Conf. Imaging Science, Systems, Technology*, Las Vegas, Nevada, USA, 2002, pp. 439—445.

[5] G. Scott *et al.*, "Entropy-balanced bitmap tree for shape-based object retrieval from large-scale satellite imagery databases," *IEEE Trans. Geosci. and Remote Sens.*, vol. 49, no. 5, pp. 1603 — 1616, May, 2011.

[6] 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.

[7] M. Ferecatu and N. Boujemaa, "Interactive remote-sensing image retrieval using active relevance feedback," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 4, pp. 818—826, Apr., 2007.

[8] Y. Li and T. Bretschneider, "Semantics-based satellite image retrieval using low-level features," in *Proc. IEEE Int. Geoscience Remote Sensing Symp.*, Anchorage, AK, USA, 2004, vol. 7, pp. 4406—4409.

[9] Y. Hongyu, L. Bicheng, and C. Wen, "Remote sensing imagery retrieval based-on Gabor texture feature classification," in *Proc. Int. Conf. Signal Processing*, Beijing, China, 2004, pp. 733—736.

[10] B. S. Manjunath and W. Y. Ma, "Texture features for browsing and retrieval of image data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 8, pp. 837—842, Aug., 1996.

[11] S. Newsam *et al.*, "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.

[12] 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.

[13] T. Ojala, M. Pietikainen and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul., 2002.

[14] B. Demir and L. Bruzzone, "Hashing-based scalable remote sensing image search and retrieval in large archives," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 2, pp. 892–904, Feb., 2016.

[15] E. Aptoula, "Remote sensing image retrieval with global morphological texture descriptors," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 3023—3034, 2014.

[16] Y. Yang, and S. Newsam, "Geographic image retrieval using local invariant features," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 2, pp. 818-832, 2013.

[17] M. Brown and S. Ssstrunk, "Multi-spectral SIFT for scene category recognition," in *IEEE Conf. Computer Vision and Pattern Recognition*, Colorado Springs, CO, USA, 2011, pp. 177-184.

[18] K. van de Sande, T. Gevers, and C. Snoek, "Evaluation of color descriptors for object and scene recognition," in *IEEE Conf. Computer Vision and Pattern Recognition*, Anchorage, AK, USA, 2008.

[19] F. S. Khan, J. Van De Weijer and M. Vanrell, "Top-down color attention for object recognition," in *IEEE 12th Int. Conf.Computer Vision*, Kyoto, Japan, 2009.

[20] R. H. Luke, J. M. Keller and J. Chamorro-Martinez, "Extending the scale invariant feature transform descriptor into the color domain," in *Proc. of the ICGST Int. Journal Graphics, Vision, Image Processing, GVIP 8*, 2008, pp. 35—43.

[21] C. Wengert, M. Douze and H. Jgou, "Bag-of-colors for improved image search," in *Proc. 19th ACM Int. Conf. Multimedia. ACM*, Scottsdale, AZ, USA, 2011.

[22] G. Shakhnarovich, T. Darrell and P. Indyk, *Nearest-Neighbor Methods in Learning and Vision: Theory and Practice*, The MIT Press, 2006.

[23] B. Demir and L. Bruzzone, "A Novel Active Learning Method in Relevance Feedback for Content Based Remote Sensing Image Retrieval", *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no.5 , pp. 2323—2334, 2015.

[24] M.-L. Zhang and Z.-H Zhou "Ml-knn: A lazy learning approach to multi-label learning," *Pattern Recognition*, vol. 40, no. 7, pp. 2038–2048, 2007.

[25] M. R. Boutell *et al.* "Ml-knn: Learning multi-label scene classification," *Pattern Recognition*, vol. 37, no. 9, pp. 1757–1771, 2004.

[26] G. Nasierding and A. Z. Kouzani, "Empirical study of multi-label classification methods for image annotation and retrieval," in *Int. Conf. Digital Image Computing: Techniques Applications*, 2010, pp. 617622.

[27] J. Wright *et al.*, "Robust face recognition via sparse representation." *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210—227, 2009.

[28] G. Qiu, "Indexing chromatic and achromatic patterns for content-based colour image retrieval," *Pattern Recognition*, vol. 35, no. 8, pp. 1675—1686, 2002.

[29] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Computer Vision*, vol. 60, no. 2, pp. 91—110, 2004.

[30] Y. C. Pati, R. Rezaiifar and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," *IEEE Conf. Rec. 27th Asilomar Conf. Signals, Systems and Computers*, 1993.

[31] Y. Yang and S. Newsam, "Bag-Of-Visual-Words and Spatial Extensions for Land-Use Classification," *ACM SIGSPATIAL Int. Conf. on Advances in Geographic Information Systems (ACM GIS)*, 2010.

[32] F.Omruuzun, B. Demir , L. Bruzzone, Y. Y. Cetin, "Content based hyperspectral image retrieval using bag of endmembers image descriptors" *8th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*, US, 2016.

[33] S. Cui, G. Schwarz, and M. Datcu, "Remote Sensing Image Classification: No Features, No Clustering" *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 11, pp. 51585170, 2015.

[34] F.-A. Georgescu, C. Vaduva, D. Raducanu, and M. Datcu, "Feature Extraction for Patch-Based Classification of Multispectral Earth Observation Images," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 6, pp. 865869, 2016.

[35] Q. Liu et al., "Learning Multiscale Deep Features for High-Resolution Satellite Image Scene Classification", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 1, pp. 117-126, 2018