Akshara Preethy Byju, Begüm Demir, Lorenzo Bruzzone

A Progressive Content-Based Image Retrieval in JPEG 2000 Compressed Remote Sensing Archives

Preethy Byju, A., Demir, B., & Bruzzone, L. (2020). A Progressive Content-Based Image Retrieval in JPEG 2000 Compressed Remote Sensing Archives. IEEE Transactions on Geoscience and Remote Sensing, 58 (8), 5739–5751. https://doi.org/10.1109/tgrs.2020.2969374

A Progressive Content Based Image Retrieval in JPEG 2000 Compressed Remote Sensing Archives

Akshara Preethy Byju[1], Begüm Demir[2] and Lorenzo Bruzzone[1]

[1]Dept. of Information Engineering and Computer Science, University of Trento, Trento, Italy
[2]Faculty of Electrical Engineering and Computer Science,Technische Universität Berlin, Berlin, Germany

***Abstract*—Due to the dramatically increased volume of remote sensing (RS) image archives, images are usually stored in compressed format to reduce the storage size. Existing content-based RS image retrieval (CBIR) systems require as input fully decoded images, thus resulting in a computationally demanding task in the case of large-scale CBIR problems. To overcome this limitation, in this paper we present a novel CBIR system that achieves a coarse to fine progressive RS image description and retrieval in the partially decoded JPEG 2000 compressed domain. The proposed system initially: i) decodes the code-blocks associated only to the coarse wavelet resolution, and ii) discards the most irrelevant images to the query image based on the similarities computed on the coarse resolution wavelet features of the query and archive images. Then, the code-blocks associated to the sub-sequent resolution of the remaining images are decoded and the most irrelevant images are discarded by computing similarities considering the image features associated to both resolutions. This is achieved by using the pyramid match kernel similarity measure that assigns higher weights to the features associated to the finer wavelet resolution than to those related to the coarse wavelet resolution. These processes are iterated until the codestreams associated to the highest wavelet resolution are decoded. Then, the final retrieval is performed on a very small set of completely decoded images. Experimental results obtained on two benchmark archives of aerial images point out that the proposed system is much faster while providing a similar retrieval accuracy than the standard CBIR systems.**

***Index Terms*—content based image retrieval, progressive image retrieval, JPEG 2000, compressed image domain, pyramid match kernel, remote sensing**

I. Introduction

DUE to the rapidly growing remote sensing (RS) image archives, content-based image retrieval (CBIR) systems with high accuracy and fast search speed are becoming an important tool in RS. Given a query image, CBIR systems aim at retrieving the most similar images ranked in terms of relevance. To achieve this goal, existing CBIR systems consist of two main modules [1]: i) image description module, which aims at characterizing the content of the images with discriminative and descriptive features, and ii) image retrieval module, which aims at matching the descriptor of the query image with those of archive images with high accuracy and returning the most similar images to the query image in a computationally efficient manner.

The performance of CBIR systems strongly depends on the capability of the RS image descriptors to accurately characterize the content of the images. Several RS image descriptors are presented in the RS literature. In [2], descriptors extracted by the scale invariant feature transform (SIFT) and their bag-of-visual-words representations have been presented, while bag-of-morphological-words representations of local morphological texture descriptors are introduced in the context of CBIR in [3]. Local binary patterns (LBPs) that represent the relationship of each pattern (i.e., pixel) in a given image with its neighbors located on a circle around that pixel by a binary code are found effective in RS. In [4], a comparative study is presented that analyzes and compares advanced LBP variants in the RS CBIR domain. An image description method that characterizes both spatial and spectral information content of high dimensional RS images is presented in [1]. This method models the spectral content by three different spectral descriptors that are: the raw pixel values, the simple bag of spectral values descriptors and the extended bag of spectral values descriptors. To model the spatial content of RS images the SIFT-based bag of visual words approach is exploited in [2]. In [5, 6] methods that model images by graphs are presented, where the nodes model region properties and the edges represent the spatial relationships among the regions. Hashing methods that represent the images with binary codes in a low-dimensional hamming space have recently received great attention in RS [7]. Deep learning methods have been recently used in the context of hashing due to the high capability of deep networks (e.g., Convolutional Neural Network) to model high-level semantic content of RS images [8, 9]. A deep hashing neural network to address CBIR in RS is introduced in [10]. This network learns to generate semantically efficient deep image features and binary hash codes by considering the cross-entropy loss. A metric learning based deep hashing network that learns a semantic-based metric space, while simultaneously producing binary hash codes for fast and accurate retrieval of RS images is presented in [11].

Once image descriptors are obtained, one can proceed with the image retrieval task. One of the simplest approach to perform RS image retrieval is to use the k-nearest neighbor (k-nn), which computes the similarity between the query image and all archive images and returns the k images most similar to the query. RS image retrieval can be also modeled as a binary-classification problem: one class includes images relevant to the query image, and the other class consists of irrelevant images. Such an approach requires the availability of annotated training images by single high-level land-use category labels (which are associated to the semantic content of the image). In this case, any binary classifier can be used in the context

of CBIR. Images can be also annotated by low-level land-cover class labels (i.e., multi-labels). To efficiently characterize images with multi-labels, multi-label image retrieval methods should be used. As an example, a sparse reconstruction-based RS image retrieval method is presented in [1]. This method considers a measure of label likelihood in the framework of sparse reconstruction-based classifiers and generalizes the standard sparse classifier to the case of both single label and multi-label RS image retrieval problems. If the RS images are represented by graphs as presented in [5, 6], image retrieval can be achieved by using graph matching techniques. As an example, in [5] image similarity is estimated using an inexact graph matching strategy, which jointly exploits a subgraph isomorphism algorithm and a spectral embedding algorithm.

All the above-mentioned image description and retrieval methods are potentially effective for RS CBIR. However RS images are usually stored in compressed format in the archives to reduce their storage size. Thus, image decoding is required before applying any image description method and thus retrieval algorithm. This is computationally-demanding and impractical in the case of large-scale RS image retrieval problems. To address this issue, in this paper we present a novel CBIR system that aims at retrieving images by minimizing the amount of image decompression without reducing the retrieval accuracy. We assume that the images in the archives are JPEG 2000 compressed, since this compression approach is widely used in the current operational RS data storage systems. The novelty of the proposed CBIR system consists in the design and development of a novel coarse to fine progressive image retrieval strategy that operates on partially decompressed images. Thus it does not require a full decoding of images for feature extraction and image retrieval. The proposed system is completely unsupervised and can be exploited by using any image descriptor that can accurately describe the wavelet coefficients. Experiments carried out on two benchmark archives demonstrate the effectiveness of the proposed progressive CBIR system in terms of the retrieval accuracy and search time.

The remaining part of this paper is organized as follows. Section II presents an overview of the JPEG 2000 algorithm and summarizes existing image retrieval methods in JPEG 2000 compressed image archives. Section III introduces the proposed coarse to fine progressive image retrieval system. Section IV presents the benchmark archive used in the experiments, while Section V illustrates the experimental results. Section VI draws the conclusion of this work.

II. RELATED WORKS

A. Overview of the JPEG 2000 algorithm

Several compression algorithms are introduced in the RS literature (e.g., Differential Pulse Code Modulation (DPCM), Adaptive DPCM, Joint Photographic Experts Group (JPEG), lossy and lossless JPEG and JPEG 2000) [12, 13]. Among these algorithms, JPEG 2000 has become very popular due to its inherent multiresolution paradigm, scalability and good compression ratio. As an example, Sentinel-2 multispectral images are compressed by applying the JPEG 2000 algorithm to each spectral band independently from each other.

Fig. 1 shows a general block scheme of JPEG 2000 compression algorithm. The key elements of the JPEG 2000 compression algorithm are [12, 14, 15]: i) 2D Discrete Wavelet Transform (2D DWT); ii) Quantization; and iii) Entropy Block Coding with optimized Truncation (EBCOT). Initially, each band of the given input image is divided into small rectangular blocks called *tiles*. Each *tile* obtained from several spectral bands of an image is compressed separately. If the size of the image is very large, dividing an image into small *tiles* improves the processing speed during the entropy-encoding step.

In the JPEG 2000 framework, successive dyadic wavelet decomposition is performed using either (9,7) or (5,3) biorthogonal filter bank. Successive dyadic wavelet decomposition applied to each *tile* separately transforms an image into one low frequency (approximation coefficients-LL) sub-band and three high frequency sub-bands (detailed coefficients-LH, HL and HH). If there are more than one decomposition level, the lowest sub-band of the current resolution is further decomposed to the subsequent lower resolution sub-bands. For lossy compression, each of the wavelet coefficients is quantized using a particular scalar value, while for the lossless compression the quantization step is neglected. Before performing the entropy coding, each sub-band is sub-divided into non-overlapping rectangular blocks called *precinct* and each *precinct* is further sub-divided into non-overlapping blocks called *code-block*s that are represented as bit planes. Each *code-block* has size usually of 32×32 or 64×64 pixels.

EBCOT that is the entropy coding technique used in the JPEG 2000 framework is subdivided into two steps: i) *Tier-1*; and ii) *Tier-2* encoding. In *Tier-1* encoding, each code-block is entropy-coded using: i) Context Modelling; and ii) Arithmetic coding. Contextual information of bit planes of these code-blocks can be obtained from three coding passes: *significant propagation pass*, *magnitude refinement pass* and *clean up pass*. In *significant propagation pass*, the bit is encoded if it is not significant or with at least one significant neighbor. In *magnitude refinement pass*, all the bits that are significant in the previous pass are coded and finally all the bits that are not coded are encoded in the *cleanup pass*. The contextual information of these code-blocks is encoded from Most Significant Bit (MSB) to Least Significant Bit (LSB) to obtain the compressed bit stream, which is performed in the *Tier-1* coding of EBCOT.

In *Tier-2* encoding, the compressed bit streams are organized into several *packets* and *layers* based on the resolution, component, spatial area and quality. *Packet* structure contains information about a few spatially consistent subgroups of *code-block* within a particular resolution, quality or level [15]. This packet structure organization allows to access the compressed bit stream of any resolution, level or component without decoding the entire compressed image. This freedom to access information regarding any level, resolution or component without decoding the entire compressed image is often termed as *scalability*. This arrangement allows a progressive encoding as well as decoding that is utilized in our proposed CBIR system.

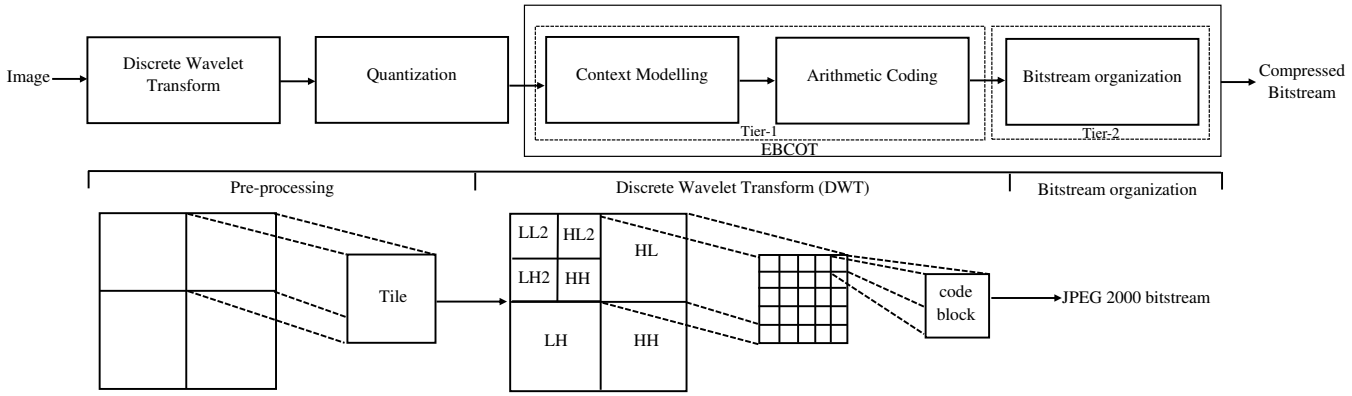


Fig. 1. A general block scheme of the JPEG 2000 compression algorithm.

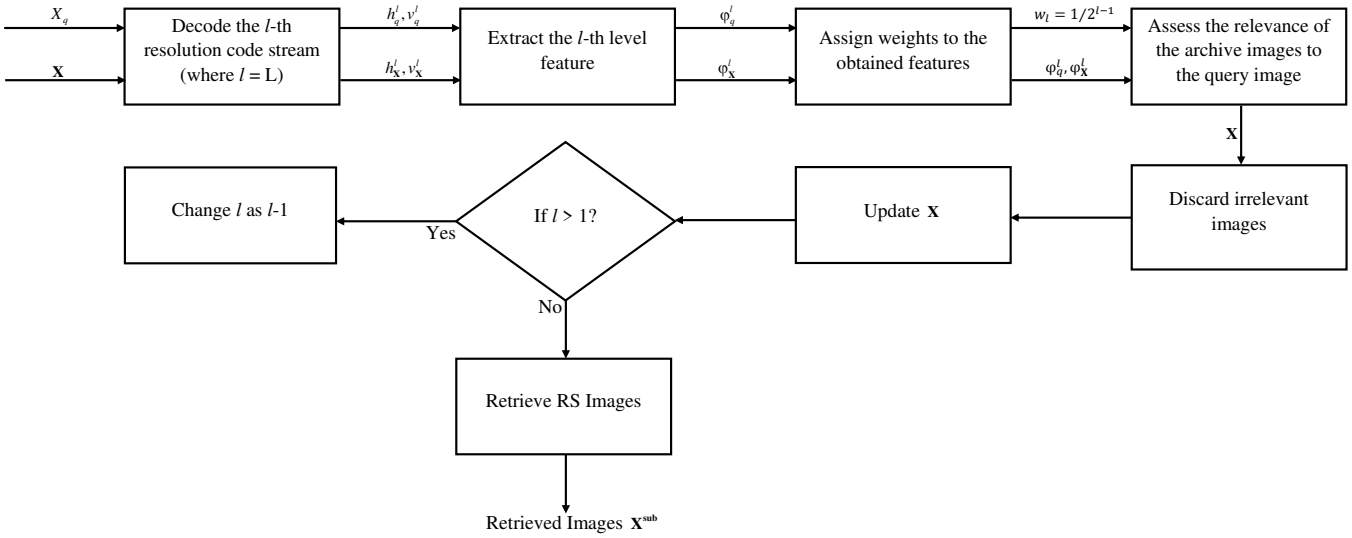


Fig. 2. Block scheme of the proposed coarse to fine progressive RS CBIR system within the JPEG 2000 framework.

B. CBIR in JPEG 2000 Compressed Image Archives

In computer vision and pattern recognition, several methods were introduced to obtain image descriptors from the JPEG 2000 compressed images. Two different types of descriptors can be considered: i) header descriptors, which describe fully compressed bit streams; and ii) wavelet descriptors, which describe partially decoded (entropy-decoded) wavelet coefficients. Header descriptors can include: a) the number of bytes used to entropy encode the image code-blocks; and b) the number of significant bit planes of a code-block [16]. To characterize images with complex semantic content for image retrieval problems, header descriptors alone is insufficient. Several wavelet descriptors have been introduced in the computer vision literature. As an example, the statistical information obtained from high and low frequency sub-bands are found effective to define image characteristics such as directionalities, shape and edges. In details, Gaussian Mixture Models (GMMs) are considered to model the high frequency wavelet sub-bands since the detail sub-bands exhibit near-Gaussian distributions [17].

In [17] spectral features are obtained from the low frequency sub-band while texture features are obtained from high frequency sub-bands by modeling the wavelet distribution using GMMs. It is worth noting that computing GMMs for images within large-scale image archives is computationally-demanding. In addition, if the number of spectral bands in an image increases, the time required to compute the features using GMM further increases. In [18], a bit-plane probability signature obtained from the joint probability distribution of the wavelet coefficients is introduced to model the texture content of the images. Moment invariants from the wavelet coefficients are introduced in [19] to define the shape information content of an image in the framework of image retrieval. In [14] the low frequency sub-band and all the high frequency sub-bands are described by using wavelet sub-band histograms for image retrieval problems. In [20], all the images in the archive are represented by a combination of energy and moment significance maps associated to each sub-band. All the above-mentioned methods require entropy decoding (i.e., partial decoding) associated to all images within the archive. This is time-demanding and computationally challenging in the case of large-scale CBIR problems.

III. PROPOSED CBIR SYSTEM

A. Problem formulation

Let $\mathbf{X} = \{X_i\}_{i=1}^{N}$ be an archive that consists of a large number of *N* JPEG 2000 compressed RS images, where X_i represents the *i*-th compressed image. Given a query image ($X_q \in \mathbf{X}$ or $X_q \notin \mathbf{X}$) the main objective of the proposed system is to retrieve a set of relevant images $X^{rel} \subset \mathbf{X}$ from the archive $\mathbf{X}$ that are semantically similar to X_q without fully decoding all the images in $\mathbf{X}$. We assume that images in the archive are compressed by using the JPEG 2000 algorithm based on *L* wavelet decomposition levels, i.e. an image X_i has one low-pass sub-band (approximation sub-band) and *3L* high-pass (horizontal, vertical and diagonal) sub-bands. Thus, the total number of sub-bands for a given image with *L* decomposition levels is 3*L*+1. When JPEG 2000 is considered, the simplest approach to perform image retrieval consists of three main steps: 1) entropy-decoding of all the images in the archive $\mathbf{X}$, 2) extraction of image descriptors, and 3) analysis of similarity and retrieval of images relevant to the query image. However, entropy-decoding all the *N* images up to *L* decomposition levels in a large-scale image archive is time-consuming and computationally challenging. To address this problem, we present a novel CBIR system that achieves a coarse to fine progressive RS image description and retrieval in the partially decoded JPEG 2000 compressed domain. Fig. 2 shows the general block scheme of the proposed system. In the following sub-sections, we initially introduce the method used for characterization of wavelet decomposition levels and then provide detailed explanation on the proposed CBIR system.

B. Characterization of Wavelet Decomposition Levels

In this paper, we characterize each wavelet decomposition level with the texture descriptors proposed in [21] as magnitudes of wavelet frame coefficients. Let $h^l_{X_i}$ and $v^l_{X_i}$ be the horizontal and vertical sub-bands of an image X_i at the *l*-th wavelet decomposition level. To define the texture descriptor $H^l_{X_i}$ of the *l*-th level, the moduli $\varphi^l_{X_i}(u,v)$ of the horizontal and vertical detail coefficients are initially estimated as follows:

$$\varphi^l_{X_i}(u,v) = \sqrt{[h^l_{X_i}(u,v) + v^l_{X_i}(u,v)]^2}, u = 1,2,...m; \quad v = 1,2,...n \tag{1}$$

where $h^l_{X_i}(u,v)$ and $v^l_{X_i}(u,v)$ represent horizontal and vertical coefficients, respectively, which are associated to the sample location (u,v) for the *l*-th sub-band of $m \times n$ size. Then the histogram $H^l_{X_i}$ of $\varphi^l_{X_i}(u,v), u = 1,2,...m; v = 1,2,...n$, which models the distribution of the moduli of wavelet sub-band coefficients, is taken as the descriptor of the *l*-th wavelet resolution. In order to estimate the histogram $H^l_{X_i}$, initially the range of possible values are defined by the minimum and maximum sample values of $\varphi^l_{X_i}$ (independently from the other wavelet decomposition levels) and the range is divided into *r* histogram intervals (i.e., *r* histogram bins). Then, the histogram $H^l_{X_i}$ is computed by counting how many of the patterns belong to each interval. Accordingly, the histogram describes a feature vector consisting of a marginal distribution of moduli of wavelet horizontal and vertical detail coefficients. Note that if a sufficient number *r* of histogram bins is defined, the histogram can represent the underlying distribution with a high precision. Thus, the histogram associated with each wavelet decomposition level of each image is capable of effectively capturing the texture content of the related image. It is worth noting that texture descriptors obtained from the lowest wavelet resolution are able to capture the global structure (coarse-scale objects) of an image, whereas the texture descriptors obtained from the higher wavelet resolutions are able to capture local detailed information (fine-scale objects). It is worth noting that the texture descriptor is a histogram-based descriptor and thus rotation and translation invariant. However, it is not scale invariant and does not explicitly model the different illumination conditions. However, the proposed system is independent from the selected descriptor and any descriptor that can accurately describe the wavelet coefficients can be used.

C. Proposed Progressive CBIR System

The proposed progressive CBIR system initially decodes the codestreams associated with the lowest wavelet resolution (i.e., *L*-th level) for all *N* images in the archive and then extracts the texture descriptor $H^L_{X_i}$ [where *l*=*L* in (1)] that models the marginal distribution of moduli of wavelet horizontal and vertical detail coefficients at the *L*-the level. Then, the similarities between the descriptors $H^L_{X_i}$ and that of the query image $H^L_{X_q}$ are measured by the Histogram Intersection (HI) kernel that is defined as [22]:

$$HI(H^L_{X_q}, H^L_{X_i}) = \sum_{i=1}^{r} min(H^r_{X_q}, H^r_{X_i}) \tag{2}$$

where *r* represents the number of histogram intervals. Then, the most dissimilar images to the query image, which are associated to the lowest similarity values, are discarded and $\mathbf{X}$ is updated. The next step starts by: i) decoding the code-blocks associated to the subsequent resolution level (i.e., *l*=*L*-1) of the remaining images in the archive and the query image; and ii) extracting their texture descriptor $H^{L-1}_{X_i}$. Accordingly, each remaining image in $\mathbf{X}$ and the query image are represented by increasingly fine descriptors associated to the first two wavelet resolutions. It is worth noting that descriptors associated to higher wavelet resolutions are capable of modeling more detailed information of the fine-scale objects in the images with respect to those associated to the lower wavelet resolutions. Thus, while estimating similarities between the query image and remaining images, we give higher weight values w_{L-1} to the descriptors associated to the (*L*-1)-th wavelet resolution than to weight values w_L of the descriptors associated to the coarsest *L*-th wavelet resolution. This is done by using the pyramid match (PM) kernel similarity measure [23], which computes the weighted sum of the all the implicit correspondences between the texture descriptors of the different wavelet decomposition levels by both considering their weights and preserving their individual distinctness at each level. The PM kernel takes a weighted sum of the number of matches (i.e., the number of samples that fall into the same histogram interval)

that occur at each level of resolution, by assigning higher weights to the matches found at higher resolution with respect to those found at coarser resolutions. The PM kernel is defined as [23]:

$$PM(H^l_{X_q}, H^l_{X_i}) = \sum_{m=l}^{L} w_m N_m, with\ l < L \qquad (3)$$

where, as suggested in [23], $w_m = 1/2^{m-1}$ and N_m shows the implicit partial correspondence between any two successive wavelet decomposition levels. Note that the number of matches found at level *L* can also include all the matches found at the finer level (*L*-1). Thus, the number of new matches found at level *L* is given by:

$$N_m = HI(H^m_{X_q}, H^m_{X_i}) - HI(H^{m-1}_{X_q}, H^{m-1}_{X_i}) \qquad (4)$$

where *m* = 1,2,...L denotes the wavelet decomposition level. It is worth noting that the PM kernel similarity measure is presented in [23] to assess the implicit partial matching correspondences between two multiresolution histograms to achieve a discriminative classification of variable feature sets. In this paper, we exploit it for estimating the similarity among the image descriptors that are associated to different wavelet decomposition levels.

After estimating the PM similarities, the most irrelevant images are discarded and **X** is updated. Then next step starts by decoding the code-blocks associated to the subsequent resolution (i.e., *l*=*L*-2) of the remaining images in **X** and describing the images by increasingly fine descriptors associated to the three wavelet resolutions. The image similarities are estimated by (4) including the three descriptors with their associated weight values, and most irrelevant images are discarded. These decoding and discarding processes are iterated until the codestreams associated to the highest wavelet resolution (i.e., when *l*=1) are decoded. Then the most similar images to the query are selected. If the images in the archive are decomposed up to *L* wavelet levels, then the number of stages that discards irrelevant images in the proposed system will be *L*-1. Due to the progressive coarse to fine CBIR mechanism, the proposed system exploits a multiresolution and hierarchical feature space to accomplish a progressive RS CBIR with an optimal use of resources in terms of retrieval and decoding time. It is also worth noting that in the final retrieval of the proposed CBIR system using the fine features, any search strategy can be adopted.

IV. Dataset description and experimental setup

To evaluate the effectiveness of the proposed system, we performed several experiments on two benchmark archives. The first one is the widely used UCMERCED benchmark archive that consists of 2100 images of size 256×256 pixels selected from aerial orthoimagery with a spatial resolution of 30 cm [2]. Images are obtained from USGS National Map Urban Area Imagery collection of the following U.S. 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. To evaluate the performance of the proposed method, we considered the annotations[1] of the images with multi-labels. The total number of the multi-labels is 17 (which are: airplane; bare-soil; buildings; cars; chaparral; court; dock; field; grass; mobile-home; pavement; sand; sea; ship; tanks; trees; water), while the number of labels associated with each image varies between 1 and 7 [5]. For the example of images with their associated multi-labels the reader is referred to [5].

The second archive is the AID benchmark archive that consists of 10,000 aerial images of size 600×600 pixels with spatial resolution variable between 0.5 m. and 8 m. To assess the effectiveness of the proposed system, we considered the annotations of the images with single labels. The total number of single labels is 30 (i.e., airport, bare land, baseball field, beach, bridge, center, church, commercial, dense residential, desert, farmland, forest, industrial, meadow, medium residential, mountain, park, parking, playground, pond, port, railway station, resort, river, school, sparse residential, square, stadium, storage tanks and viaduct). For examples of images and their labels the reader is referred to [24].

To assess the effectiveness of the proposed system, the images of both archives were initially compressed by the JPEG 2000 algorithm by using 3 wavelet decomposition levels (i.e., $L = 3$). It is worth noting that since the size of the code-block used to obtain the JPEG 2000 compressed codestream must not be less than 32×32 pixels to obtain relevant information from the compressed images, in both archives it is not possible to use $L > 3$. Each sub-band is represented by a 24-dimensional feature descriptor. After decoding the codestreams associated to the lowest decomposition level (*l*=3), T_1 of the most irrelevant images are discarded, where T_1 represents the percentage of discarded images. Then, T_2 of the most irrelevant images are discarded after decoding the second lowest decomposition level (*l*=2), where T_2 represents the percentage of discarded images at the second level. Finally, the image retrieval is performed based on the *k*-nearest neighbor (*k*-nn) search strategy by using jointly the features obtained from the highest wavelet decomposition level and the previous levels from the remaining subset of relevant images.

Results of each system for the UCMERCED archive are provided in terms of: i) average recall, ii) average precision, and iii) average computational time obtained in 2100 trials performed with 2100 selected query images from the archive. For the details on how the recall and precision are estimated in the framework of multi-label image search and retrieval problems, the reader is referred to [5]. The results obtained for the AID archive are provided in terms of (i) average precision, and (ii) average computational time associated to 10,000 trials with 10,000 selected query images from the archive. Note that while we estimate the average precision and recall for multi-label image retrieval, for the single label case, average precision and recall reduce to the same performance measures as that of the multi-label image retrieval. Thus, we report only the average precision values for the single label retrieval experiments. The retrieval performance for both archives was

[1] Annotations are available at http://bigearth.eu/datasets.html.

TABLE I
COMPARISON OF THE PERFORMANCE FOR DIFFERENT DESCRIPTORS (UCMERCED ARCHIVE).

Descriptors	Average Precision (%)	Average Recall (%)	Feature Extraction time (seconds)
EES [18]	59.80	61.73	7.11
LBP [25]	47.76	47.79	9.08
GLCM [26]	61.84	64.01	34.28
EES and LBP	59.24	60.84	20.97
LEH [27]	59.80	62.18	11.07
DMHV	**68.18**	**70.87**	**29.08**

assessed on the top-20 retrieved images. All the experiments are implemented via MATLAB® on a standard PC with Intel®Xeon®CPU i3-6100 @ 3.40GHz, 16GB RAM.

V. EXPERIMENTAL RESULTS

We carried out several experiments in order to: 1) compare the effectiveness of the considered descriptor that models the distribution of moduli of the horizontal and vertical detail coefficients (called as the DMHV descriptor hereafter) with respect to the popular descriptors that model the wavelet coefficients; 2) conduct a performance analysis with respect to varying values of T_1 and T_2 of discarded images after decoding codestreams associated with the first two wavelet decomposition levels; and 3) evaluate and compare the effectiveness of the proposed system with respect to (i) a standard-CBIR system using SIFT features obtained from fully-decoded images; (ii) a standard-CBIR system using DMHV descriptors without coarse to fine strategy.

A. Comparison of the image descriptors in the compressed domain

In the first set of trials, we analyze and compare the effectiveness of the proposed DMHV descriptor with the widely used descriptors adapted with wavelet coefficients in the literature. The selected descriptors are: 1) the extended energy signature (EES) [18]; 2) the local binary pattern (LBP) [25]; 3) the gray level co-occurrence matrix (GLCM) based measure [26]; 4) the joint use of the EES and the LBP; and 5) the local energy histogram (LEH) [27]. To have a fair comparison, we applied the DMHV descriptor to the entropy decoding of all wavelet decomposition levels (i.e., the coarse to fine retrieval strategy is not considered). Tables I and II show the results obtained for the UCMERCED and AID archives, respectively. From the tables, one can observe that the DMHV descriptor provides the highest accuracy for both archives. This is achieved at the cost of increasing the required computational time. The GLCM-based descriptor provides the second best performance. In detail, the DMHV descriptor results in an improvement of almost 6.34% and 6.86% in average precision and average recall, respectively for the UCMERCED archive when compared to the GLCM based descriptor with slightly higher computational time.

Fig. 3 and 4 show an example of images retrieved from the UCMERCED and AID archives, respectively, by considering all the above-mentioned descriptors. In Fig. 3 the query image includes *bare soil*, *buildings*, *cars*, *pavement* and *trees*. The retrieval order and the multi-labels associated with each image are given above and below the related image, respectively. By analyzing the figure one can observe that all the images retrieved by using the proposed DMHV descriptor [see Fig. 3(g)] contain almost all the class labels included in the query image. On the contrary, the images retrieved by using the other descriptors mostly contain only one or two of the class labels [see Fig. 3(b-f)]. In Fig. 4 the selected query image is from *dense residential* category and the retrieved images with their associated single labels are provided below the related image. By analyzing the figure one can observe that all the images retrieved by using the DMHV descriptor [see Fig. 4(g)] belong to the *dense residential* category. When the other descriptors are used, some irrelevant images associated with category labels *airport*, *baseball field* and *sparse residential* are retrieved. By a visual analysis of all these results, we observe that the DMHV descriptor accurately models the content associated with each query image, resulting in retrieval of the visually most similar images from the archive.

B. Performance of the proposed system versus T_1 and T_2 values

In this subsection, we analyze the performance of the proposed progressive-CBIR system with respect to the parameters T_1 and T_2. Tables III and IV report the performance measures obtained after performing pyramid match kernel similarity measure using features obtained from each wavelet decomposition level for UCMERCED and AID archives, respectively. By analyzing the tables, one can see that there is a significant improvement in the performance measures when hierarchical weights are assigned to the progressively obtained coarse to fine features in the proposed image retrieval system in the compressed domain. The implicit correspondence between the feature sets obtained between any two wavelet decomposition levels adds more discriminant texture information, which is utilized to discard irrelevant images to the query image at a very early stage. In our experiments the value of the parameter T_1 that represents the percentage of images discarded at the first level is varied in the range between 0% and 100% with step-size increment of 5%.

Fig. 5 and 6 show the performance of the proposed RS CBIR system in terms of precision and recall versus the varying values of T_1 and T_2 for the UCMERCED and the AID archives, respectively. By analyzing Fig. 5 (UCMERCED archive), one can notice that there is no change in the performance measures when the value of T_1 varies between 0% and 90%. This shows that the compressed domain texture features obtained at a very coarse level are able to efficiently characterize the images in the archive. In other words, we can conclude that the DMHV descriptor are able to efficiently discriminate 90% of the images in the archive using only the coarser features. We can see a continuous decrease in the performance metrics when T_1>90% of the images are discarded at a very early stage using coarse features because of discarding relevant images in the initial stage. This occurs because descriptors

TABLE II
COMPARISON OF THE PERFORMANCE FOR DIFFERENT DESCRIPTORS (AID ARCHIVE).

Performance Metric	EES [18]	LBP [25]	GLCM [26]	EES and LBP	LEH [27]	**DMHV**
Average Precision (%)	51.34	49.97	52.97	51.17	50.57	**59.97**
Feature Extraction Time (seconds)	23.14	30.61	73.91	65.73	40.57	**75.59**

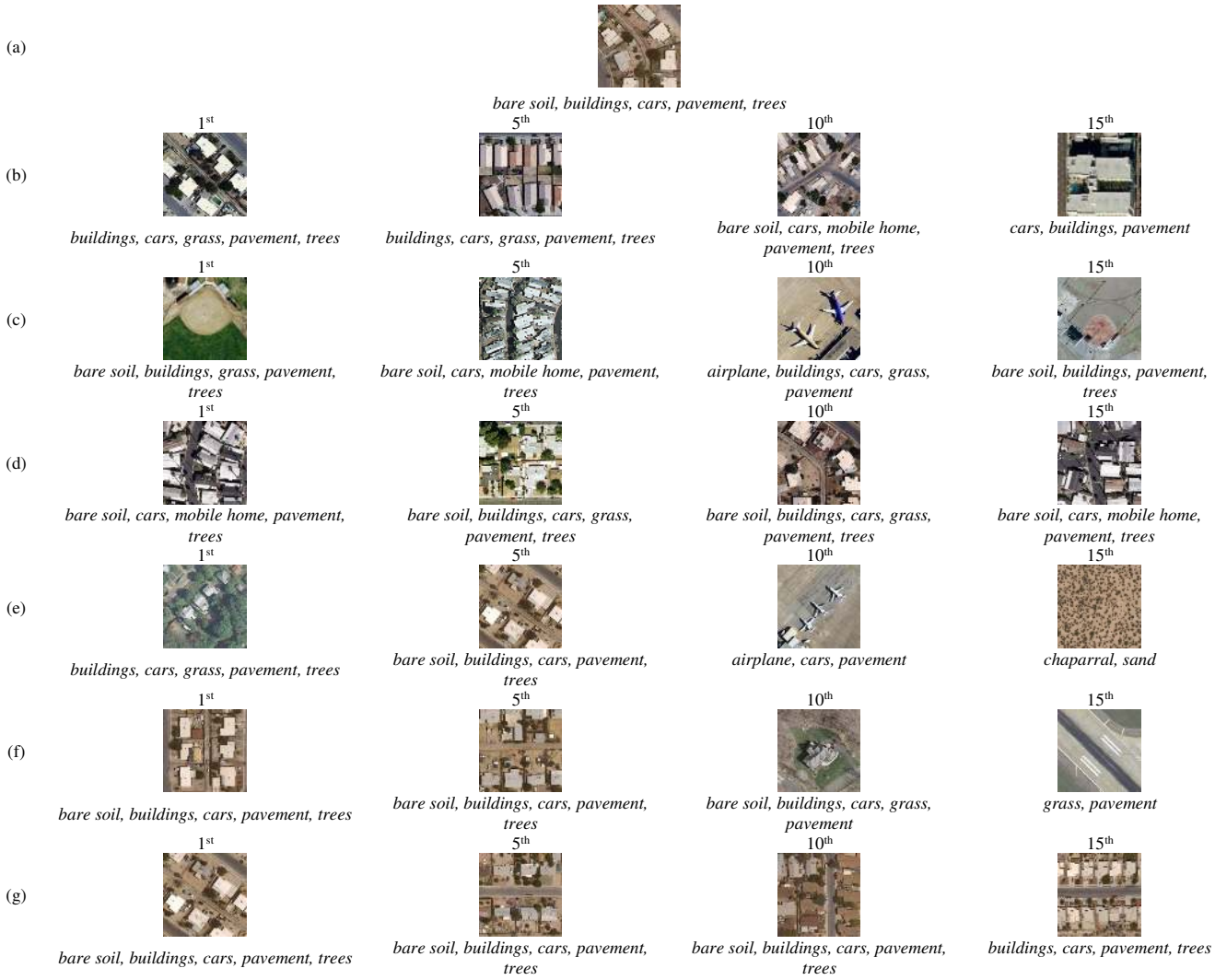


Fig. 3. Example of (a) query image; and retrieved images by using (b) the EES descriptor, (c) the LBP descriptor, (d) the GLCM descriptor, (e) a combination of the EES and the LBP descriptors, (f) the LEH descriptor and (g) the DMHV descriptor (UCMERCED archive).

TABLE III
AVERAGE PRECISION AND RECALL FOR THE PROPOSED PROGRESSIVE COARSE TO FINE RS CBIR SYSTEM AT EACH LEVEL WHEN T_1=25% FOR THE UCMERCED ARCHIVE.

Decomposition Levels	Average Precision (%)	Average Recall (%)
Level 1 **(Coarsest Feature)**	65.04	67.09
Level 2 (Fine Feature)	67.76	67.79
Level 3 **(Finest Feature)**	68.28	70.94

TABLE IV
AVERAGE PRECISION FOR THE PROPOSED PROGRESSIVE COARSE TO FINE RS CBIR SYSTEM AT EACH LEVEL WHEN T_1=25% (AID ARCHIVE).

Decomposition Levels	Average Precision (%)
Level 1 **(Coarsest Feature)**	55.67
Level 2 (Fine Feature)	57.74
Level 3 **(Finest Feature)**	60.12

obtained from the coarser resolution are able to characterize relevant images with respect to the query image. From the Fig. 6 (AID archive) we observed that there is no change in the precision values when the value of T_1 varies between 0% and 75%. This demonstrates the ability of the features obtained from the coarser level to efficiently characterize images having varying spatial resolution in the AID archive. Thus, the value of T_1 should be selected on the basis of a trade-off analysis between computational complexity and performance of the

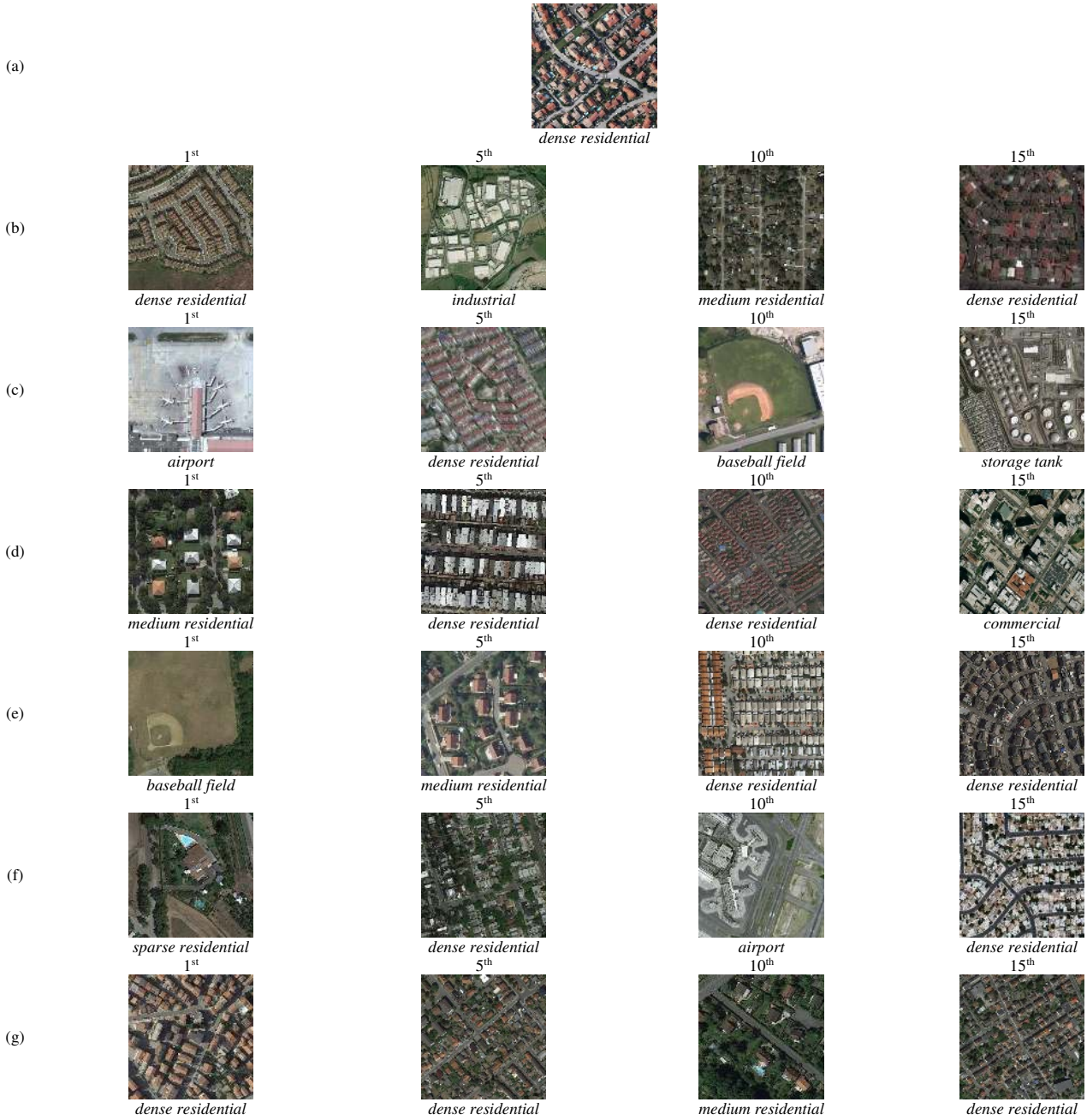
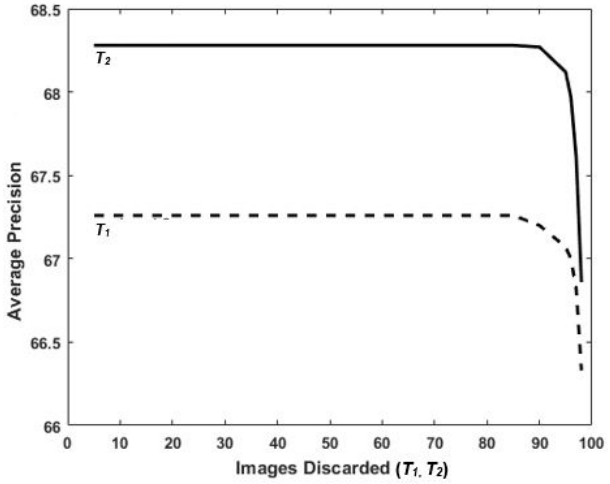


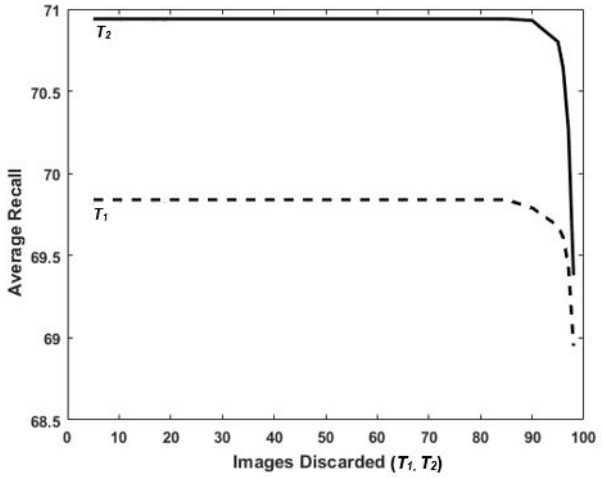
Fig. 4. Example of (a) query image; and retrieved images by using (b) the EES descriptor, (c) the LBP descriptor, (d) the GLCM descriptor, (e) a combination of the EES and the LBP descriptors, (f) the LEH descriptor, and (g) the DMHV descriptor (AID archive).

final retrieved images. On the basis of these results, we fixed the value of T_1 as 25%.

To further investigate the performance of the proposed system, we analyzed the characteristics of the images that are discarded using the coarse features and second lowest fine features. Fig. 7 and 8 show an example of the images discarded using the coarsest features and the second lowest fine features for both the archives. In detail, Fig. 7 demonstrates the discarded images when a query image that contains *bare soil, buildings, cars, pavement, trees* is selected from the UCMERCED archive. From the analysis, one may observe that using only the coarsest level features, one can discard highly irrelevant images to the query image that contain labels such as dock, ship, water, forest, mobile-home. This shows that the coarse level features are enough to reject highly irrelevant images from the archive at a very early stage. This further speeds up the proposed system as only a subset of relevant images requires decoding. In the second iteration, using the fine features, the system is able to discriminate properly highly similar images with respect to the given query image. Thus,



(a)



(b)

Fig. 5. (a) Average precision and (b) average recall provided by the proposed PCF-CBIR system versus T_1 and T_2 (UCMERCED archive).

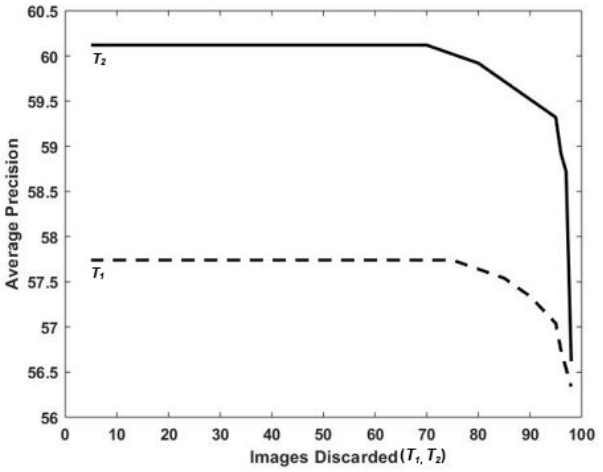


Fig. 6. Average precision provided by the proposed PCF-CBIR system versus T_1 and T_2 (AID archive).

bare soil, buildings, cars, pavement, trees

(a)

dock, ship,water | *trees* | *grass, sand, trees, water* | *mobile-home, pavement, trees*

(b)

cars, pavement, trees | *court, grass, pavement* | *bare-soil, grass, buildings, pavement* | *cars, pavement, trees*

(c)

Fig. 7. Example of (a) query image; (b) images discarded using coarsest features; and (c) images discarded using second lowest fine features (UCMERCED archive).

using the second lowest fine feature [see Fig. 7(c)] the images with more similar class label-sets such as cars, pavement, trees are discarded. We observed similar results when the AID archive is considered. Fig. 8 shows the images discarded when a query image is selected from *dense residential* category of the AID archive. By analyzing the figure, one can observe that using only the coarsest level features [see Fig. 8(b)], one can discard irrelevant images that contain labels such as *beach*, *forest*, *pond* and *center*. Using the second lowest fine features [see Fig. 8(c)], the proposed system is able to discriminate images that contain label sets such as *port*, *railway station*, *park* and *parking*.

Fig. 9 shows the behaviour of the computational time (including both decoding and feature extraction) versus the percentage of the images discarded after decoding the 2^{nd} wavelet decomposition level (for which T_1 images are discarded) and the 1^{st} wavelet decomposition level (for which T_2 images are discarded) for the UCMERCED archive. Initially, the coarse features are obtained after decoding all the *N* images in the archive. Then, T_1 of the images are discarded and the second lowest fine features are obtained for the subset of the remaining relevant images. Fig. 9-a shows the computational time required to decode and obtain features from the 2^{nd} wavelet decomposition level. From the graph, one can notice that, as the percentage of images discarded increases, the computational time required to decode and obtain features from the resulting subset of relevant images decreases. Fig. 9-b shows the required computational time to decode and obtain features from the 1^{st} wavelet decomposition level. When the value of T_1 decreases, the time taken to decode and obtain features after eliminating T_2 (which is defined as $1 - T_1$) also decreases and vice-versa. Thus, the graph is not linear and the computational time peaks when T_1= T_2=50%. From

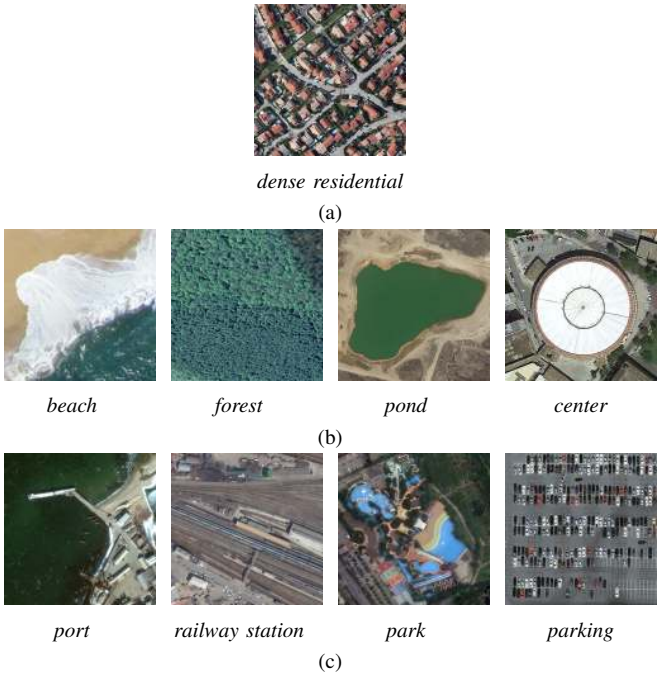


Fig. 8. Example of (a) query image; (b) images discarded using coarsest features; and (c) images discarded using second lowest fine features (AID archive).

an analysis of the peaks of the graphs, we can conclude that when we reduce the number of images that require decoding to a high wavelet decomposition level, the computational time taken by the retrieval system decreases. The same behavior is also obtained when the AID archive is used.

C. Comparison of the proposed CBIR system with the state-of-the-art systems

In this subsection, we compare the effectiveness of the proposed system (proposed progressive-CBIR) with: i) a standard-CBIR system that exploits the SIFT features obtained from fully decoded images; ii) a standard-CBIR system that exploits the DMHV descriptors without coarse to fine strategy. Tables V and VI report the results for the UCMERCED and the AID archives, respectively, along with the required decoding time and CBIR time. The decoding time is associated to the time required for decoding the codestreams, whereas the CBIR time is associated to the time taken by both the extraction of the descriptors and the retrieval of the images. It is worth noting that in the proposed progressive-CBIR system decoding of an image from the archive depends up on its relevancy in the retrieval with respect to the given query image. From the tables, one can observe that the proposed progressive-CBIR system provides higher accuracies with significantly reduced decoding and CBIR times for both archives compared to the standard-CBIR system that uses SIFT features. As an example, the proposed system outperforms the standard CBIR system by almost 3% in precision and 2% in recall for the UCMERCED archive, and almost 4% in average precision for the AID archive. The accuracies obtained by using the standard-CBIR system that exploits the same descriptor without applying the proposed coarse to fine strategy are

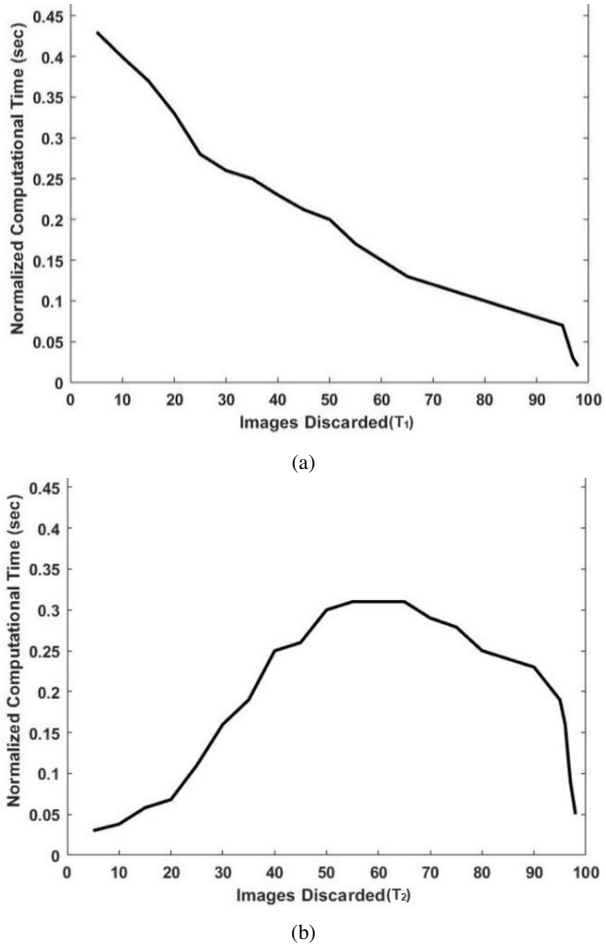


Fig. 9. Variation in computational time (including both decoding and feature extraction) versus (a) T1 and (b) T2 (UCMERCED archive).

very similar to those obtained by the proposed system for both archives. However, required decoding and CBIR times for the proposed progressive-CBIR system are almost half of the time required for the standard systems. In detail, for the UCMERCED archive the standard RS CBIR system that uses SIFT features (which requires complete decoding of the images) takes 113.65 seconds as the decoding time, whereas the proposed system takes 58.18 seconds. This shows that there is a sharp improvement in computational time (with same performance measures as the standard-CBIR system) when the image retrieval is performed in the compressed domain. All these results confirm that in the proposed system, discarding irrelevant images and adopting a progressive coarse to fine strategy shows significant improvements over the existing RS CBIR systems. Note that, as shown in the tables, discarding irrelevant images at a very early stage considerably reduces the CBIR time. Fig. 10 shows an example of results with a query image selected from the UCMERCED archive that includes six class-labels: *bare soil*, *buildings*, *cars*, *pavement* and *tree*, while Fig. 11 shows an example of results with a query image that belongs to the *baseball field* category within the AID archive. Through these examples one can see that the images retrieved from the progressive-CBIR are more relevant than those retrieved by the standard-CBIR system that uses SIFT

TABLE V
AVERAGE PRECISION AND RECALL OF THE STANDARD CBIR SYSTEM THAT USES SIFT FEATURES, STANDARD RS CBIR SYSTEM WITHOUT COARSE TO FINE STRATEGY AND THE PROPOSED PROGRESSIVE COARSE TO FINE RS CBIR SYSTEM (UCMERCED ARCHIVE).

Method	Average Precision(%)	Average Recall(%)	Decoding time (in seconds)	CBIR time (in seconds)
Standard-CBIR (SIFT features)	65.68	68.73	113.65	161.50
Standard-CBIR (DMHV descriptors without coarse to fine approach)	68.18	70.87	99.74	51.14
Proposed progressive-CBIR	**68.28**	**70.94**	**58.18**	**29.08**

TABLE VI
AVERAGE PRECISION AND RECALL OF THE STANDARD CBIR SYSTEM THAT USES SIFT FEATURES, STANDARD RS CBIR SYSTEM WITHOUT COARSE TO FINE STRATEGY AND THE PROPOSED PROGRESSIVE COARSE TO FINE RS CBIR SYSTEM (AID ARCHIVE).

Method	Average Precision(%)	Decoding time (in seconds)	CBIR time (in seconds)
Standard-CBIR (SIFT features)	55.29	259.65	271.44
Standard-CBIR (DMHV descriptors without coarse to fine approach)	59.97	234.25	141.31
Proposed progressive-CBIR	**60.12**	**127.56**	**75.59**

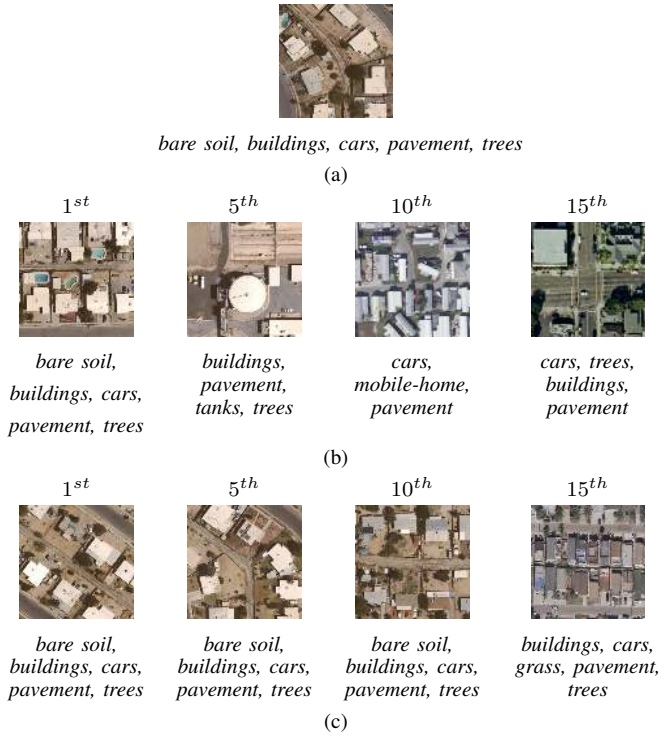


Fig. 10. Example of (a) query image; and retrieved images by using (b) the standard-CBIR system that uses SIFT features; and (c) the proposed progressive-CBIR system (UCMERCED archive).

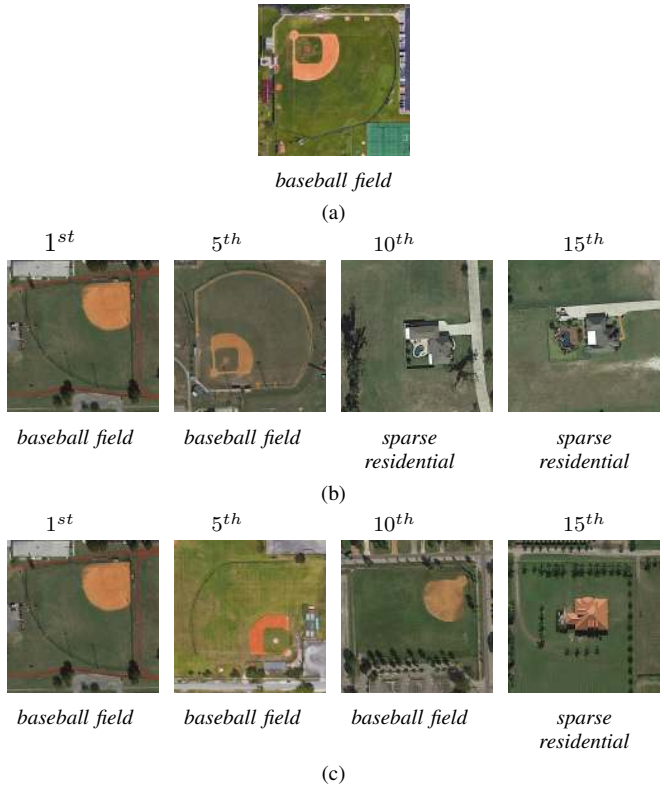


Fig. 11. Example of (a) query image; and retrieved images by using (b) the standard system without coarse to fine strategy; and (c) retrieved images by proposed progressive-CBIR system (AID archive).

features for both archives. As an example, the images retrieved using standard-CBIR system based on SIFT features does not include most of the class label sets as that of the query image for the UCMERCED archive (Fig. 10).

VI. CONCLUSION

In this paper we have introduced a novel content-based image retrieval (CBIR) system that accomplishes a coarse to fine progressive RS image description and retrieval in partially decoded JPEG 2000 compressed domain. The proposed system considers that the amount of data that needs to be entropy decoded is directly related to the relevancy of the images in the retrieval process. To reduce the time required for fully-decoding images, the proposed system initially decodes only the code-blocks associated to the lowest wavelet resolution of all images in the archive. Then, based on the similarities estimated by the histogram intersection kernel among the coarse resolution wavelet descriptors of the query image and

those of the archive images, the most irrelevant images related to the smallest similarity values are discarded. This step allows identification and elimination of the most irrelevant images at a very early stage to reduce the subsequent decoding time. The processes of code-blocks decoding and elimination of the irrelevant images (with respect to the similarities among the descriptors associated to the considered wavelet resolutions) are iterated until the codestreams associated to the highest wavelet resolution are decoded. Then, the most similar images to the query are selected. By this way, the proposed system exploits a multiresolution and hierarchical feature space representation and accomplishes a progressive RS CBIR with significantly reduced retrieval time. To characterize each resolution level, a texture descriptor that models the distribution of moduli of the horizontal and vertical detail coefficients is used. In order to evaluate the similarities among the descriptors that model different wavelet resolutions, the pyramid match kernel is exploited. The pyramid match kernel computes the weighted sum of the all the implicit correspondences between the descriptors of the different wavelet decomposition levels by considering the importance of the descriptors at different wavelet resolution levels.

Experimental results obtained on a benchmark archive show that the proposed system results in similar accuracies with respect to a standard-CBIR system (which operates on the fully decoded image domain) with significantly reduced decoding and thus retrieval time. This is due to the progressive removal of a very large amount of irrelevant images, which allows to apply the final retrieval process only to a very small set of images (which are highly relevant to the query image). We emphasize that this is a very important advantage, because the main objective of large-scale CBIR is to optimize the search and retrieval time with a minimum amount of fully decoded images. Thus, the proposed system is promising for possible operational applications due to both its general properties and also its simplicity in the implementation. Note that the archives used in the experiments are benchmarks. However, in many real applications the search is expected to be applied to much larger archives. For large scale CBIR problems, by using our system the gain in both retrieval and decoding time is expected to be increased considerably with respect to the standard-CBIR systems. As a final remark, we point out that the proposed system can be easily adapted to the CBIR problems for which images are compressed by other compression algorithms by properly defining the image description algorithm in the (partially) compressed domain. As a future development, we plan to: i) extend the validation of the proposed system to larger archives; and ii) develop deep networks that can model the images and apply retrieval in the compressed domain.

VII. ACKNOWLEDGEMENTS

This work is funded by the European Research Council (ERC) through the ERC-2017-STG BigEarth Project under grant 759764.

REFERENCES

[1] O. E. Dai, B. Demir, B. Sankur, L. Bruzzone, "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 Sens. (J-STARS)*, vol. 11, no. 7, pp. 2473--2490, 2018.

[2] 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.

[3] E. Aptoula, "Remote Sensing Image Retrieval With Global Morphological Texture Descriptors," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 5, pp. 3023–3034, 2014.

[4] I. Tekeste and B. Demir, "Advanced Local Binary Patterns for Remote Sensing Image Retrieval," in *Int. Conf. on Geoscience and Remote Sensing Symp.*, 2018, pp. 6855–6858.

[5] B. Chaudhuri, B. Demir, S. Chaudhuri, and L. Bruzzone, "Multilabel Remote Sensing Image Retrieval Using a Semisupervised Graph-Theoretic Method," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 1144–1158, 2018.

[6] B. Chaudhuri, B. Demir, S. Chaudhuri, and L. Bruzzone, "Region-Based Retrieval of Remote Sensing Images Using an Unsupervised Graph-Theoretic Approach," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 7, pp. 987–991, 2016.

[7] 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, 2016.

[8] Y. Boualleg and M. Farah, "Enhanced Interactive Remote Sensing Image Retrieval with Scene Classification Convolutional Neural Networks Model," in *IEEE Int. Conf. on Geoscience and Remote Sensing Symp.*, 2018, pp. 4748–4751.

[9] F. Ye, H. Xiao, X. Zhao, M. Dong, W. Luo, and W. Min, "Remote Sensing Image Retrieval Using Convolutional Neural Network Features And Weighted Distance," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 10, pp. 1535–1539, 2018.

[10] Y. Li, Y. Zhang, X. Huang, H. Zhu, and J. Ma, "Large-Scale Remote Sensing Image Retrieval by Deep Hashing Neural Networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 950–965, 2018.

[11] S. Roy, E. Sangineto, B. Demir, and N. Sebe, "Deep Metric and Hash-Code Learning for Content-Based Retrieval of Remote Sensing Images," in *IEEE Int. Conf. on Geoscience and Remote Sensing Symp.*, 2018, no. 1, pp. 4539–4542.

[12] S. Zhou, C. Deng, B. Zhao, Y. Xia, Q. Li, and Z. Chen, "Remote Sensing Image Compression: A Review," in *IEEE Int. Conf. on Multimedia Big Data*, 2015, pp. 406–410.

[13] G. K. Wallace, "The JPEG still picture compression standard," *IEEE Tran. on Consumer Electronics*, vol. 38, no. 1, pp. xviii–xxxiv, 1992.

[14] M. Lamard, W. Daccache, G. Cazuguel, C. Roux, and B. Cochener, "Use of a JPEG-2000 Wavelet Compression Scheme for Content-Based Ophtalmologic Retinal Images Retrieval," in *IEEE Engineering in Medicine and Biology 27th Annual Conf.*, 2005, pp. 4010–4013.

[15] C. Lui, A Study of the JPEG-2000 Image Compression Standard, PhD Thesis, University of Ontario, Ontario, Canada, 2001.

[16] A. Descampe, C. De Vleeschouwer, P. Vandergheynst, and B. Macq, "Scalable feature extraction for coarse-to-fine JPEG 2000 image classification," *IEEE Tran. on Image Processing*, vol. 20, no. 9, pp. 2636–2649, 2011.

[17] A. Teynor, M. Wolfgang, and K. Wolfgang, "Compressed Domain Image Retrieval Using JPEG2000 and Gaussian Mixture Models," in *Int. Conf. on Advances in Visual Information Systems*, 2005, pp. 132–142.

[18] M. H. Pi, C. S. Tong, S. K. Choy, and H. Zhang, "A Fast and Effective Model for Wavelet Subband Histograms and Its Application in Texture Image Retrieval," *IEEE Trans. on Image Processing*, vol. 15, no. 10, pp. 3078–3088, 2006.

[19] J. Jiang, B. F. Guo, and S. Ipson, "Shape-based image retrieval for JPEG-2000 compressed image databases," *Multimedia Tools and Applications*, vol. 29, no. 2, pp. 93–108, 2006.

[20] K. M. Au, N. F. Law, W. C. Siu, "Direct Image Retrieval in JPEG and JPEG-2000," in *IEEE Int. Conf. on Image Processing*, 2005, vol. 2000, pp. 1–4.

[21] E. De Ves, D. Acevedo, A. Ruedin, and X. Benavent, "A statistical model for magnitudes and angles of wavelet frame coefficients and its application to texture retrieval," *Pattern Recognition*, vol. 47, no. 9, pp. 2925–2939, 2014.

[22] A. Barla, F. Odone, and A. Verri, "Histogram Intersection Kernel for Image Classification," in *Int. Conf. on Image Processing*, 2003, vol. 3, pp. III–513.

[23] K. Grauman and T. Darrell, "The Pyramid Match Kernel: Efficient Learning with Sets of Features," *Journal of Machine Learning Research*, vol. 8, pp. 725–760, 2007.

[24] G. S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, and L. Zhang, "AID: A Benchmark Dataset for Performance Evaluation of Aerial Scene Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 39653981, 2017.

[25] T. Ojala, M. Pietikinen, and T. Menp, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Tran. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[26] R. M. Haralick, K. Shanmugham, and I. Dinstein, "Textural Features for Image Classification," *IEEE Trans. on Systems, Man and Cyberbectics*, vol. 3, no. 6, pp. 610–621, 1973.

[27] Y. Dong and J. Ma, "Wavelet-Based Image Texture Classification Using Local Energy Histograms," *IEEE Signal Processing Letters*, vol. 18, no. 4, pp. 247-250, 2011.

[28] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int. Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.