


Article

A Benchmark Dataset for Performance Evaluation of Multi-Label Remote Sensing Image Retrieval

Zhenfeng Shao, Ke Yang and Weixun Zhou * 

State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China; shaozhenfeng@whu.edu.cn (Z.S.); xiaoke1993@whu.edu.cn (K.Y.)

* Correspondence: weixunzhou1990@whu.edu.cn

Received: 11 May 2018; Accepted: 14 June 2018; Published: 16 June 2018



Abstract: Benchmark datasets are essential for developing and evaluating remote sensing image retrieval (RSIR) approaches. However, most of the existing datasets are single-labeled, with each image in these datasets being annotated by a single label representing the most significant semantic content of the image. This is sufficient for simple problems, such as distinguishing between a building and a beach, but multiple labels are required for more complex problems, such as RSIR. This motivated us to present a new benchmark dataset termed “MLRSIR” that was labeled from an existing single-labeled remote sensing archive. MLRSIR contained a total of 17 classes, and each image had at least one of 17 pre-defined labels. We evaluated the performance of RSIR methods ranging from traditional handcrafted feature-based methods to deep-learning-based ones on MLRSIR. More specifically, we compared the performances of RSIR methods from both single-label and multi-label perspectives. These results presented the advantages of multiple labels over single labels for interpreting complex remote sensing images, and serve as a baseline for future research on multi-label RSIR.

Keywords: remote sensing image retrieval (RSIR); multi-label benchmark dataset; multi-label image retrieval; single-label image retrieval; handcrafted features; convolutional neural networks

1. Introduction

With the rapid development of remote sensing technology, a considerable volume of remote sensing data becomes available on a daily basis. The huge amount of data has provided the literature with new opportunities for various remote sensing applications; however, it also results in the significant challenge of searching the large remote sensing archives.

Content-based image retrieval (CBIR) aims to find the images of interest from a large-scale image archive, which is a useful solution to solve this problem. Content-based remote sensing image retrieval is a specific application of CBIR in remote sensing field. Typically, an RSIR system has two main parts, feature extraction and a similarity measure, but the remote sensing community has been focused only on developing powerful features, since the performance depends greatly on the effectiveness of the extracted features.

There are a number of conventional RSIR approaches that are available and have been evaluated on the existing benchmark datasets, providing baseline results for RSIR research. However, these approaches assume that the query image, and those images to be retrieved, are single-labeled since the images are annotated by single labels associated with the main semantic content of the images. It is reasonable to make such an assumption, which is often sufficient for some particular remote sensing applications, but tends to be impossible for more complex applications. For example, single labels (broad class) are sufficient to distinguish image categories like “building” and “grass land”, but multiple labels (primitive class) are needed to distinguish between image categories like “dense

residential” and “medium residential” since they are pretty similar and the main differences lie in the density of buildings. From the perspective of RSIR, multiple labels are able to narrow down the semantic gap between low-level features and high-level semantic concepts present in remote sensing images and further improve the retrieval performance. However, the lack of a multi-label benchmark datasets has restricted the development of RSIR research. In this paper, we first introduce a new multi-label dataset, named MLRSIR, which provides the remote sensing community with a benchmark dataset to develop novel approaches for multi-label RSIR. We then provide a review of traditional single-label RSIR, as well as the multi-label RSIR approaches, ranging from handcrafted feature-based methods to deep learning feature-based ones.

The main contributions of this paper are as follows:

- We construct a multi-label remote sensing benchmark dataset, MLRSIR, for multi-label RSIR. MLRSIR is a publicly available dataset, which is a multi-labeled dataset in contrast to the existing single-labeled RSIR datasets.
- We provide a brief review of the state-of-the-art methods for single-label and multi-label RSIR.
- We compare the single-label and multi-label retrieval methods on MLRSIR, including traditional handcrafted features and deep learning features. This indicates the advantages of multi-label over single-label for complex remote sensing applications like RSIR and provides the literature with baseline results for future research on multi-label RSIR.

The rest of this paper is organized as follows. We provide a brief review of the state-of-the-art single-label and multi-label retrieval methods for RSIR in Section 2. Section 3 introduces our multi-label benchmark dataset and the multi-label RSIR methods evaluated on the dataset including handcrafted features and deep learning features. The results and comparisons are shown in Section 4. We draw some conclusions in Section 5.

2. Remote Sensing Image Retrieval Methods

RSIR is a useful technique for the fast retrieval of images of interest from a large-scale remote sensing archive. In this section, we introduce the state-of-the-art RSIR methods including handcrafted features and deep-learning-based ones from the perspective of single-label and multi-label RSIR, respectively.

2.1. Single-Label RSIR

For single-label RSIR methods, the query image and the images to be retrieved are labeled by a single, broad class label. Early single-label RSIR methods extracted handcrafted low-level features to describe the semantic content of remote sensing images, which can be either global or local features. Color (spectral) features [1], texture features [2–4], and shape features [5] are commonly used global features extracted from the whole image, while local features like Scale Invariant Feature Transform (SIFT) [6], are extracted from image patches of interest.

Color and texture features are used more widely for RSIR compared to shape features. Remote sensing images usually have multiple spectral bands (e.g., multi-spectral imagery) and even hundreds of bands (e.g., hyper-spectral imagery); therefore, spectral features are significant for remote sensing images. Bosilj et al. employed pattern spectral features for the first time in a dense strategy and explored both global and local pattern spectral features for image retrieval [1]. The results indicated that the morphology-based spectral features achieved the best performance. Color features, however, do not work sometimes due to the phenomena where the same object/class varies in spectra, or the same spectra are shared between different objects/classes. Texture features have therefore been used to capture spatial variation of pixel intensity of images, and has achieved great performance in many tasks, including RSIR. Aptoula developed multi-scale texture descriptors, the circular covariance histogram, and the rotation-invariant point triplets for image retrieval, and exploited the Fourier power spectrum as a couple of new descriptors [2]. Bouteldja et al. proposed a rotation and scale invariant

representation of the texture feature vectors by calculating the statistical measures of decomposed image sub-bands [3]. However, most of these texture features are extracted from grayscale images, and thus the rich color information is ignored. Shao et al. proposed an improved texture descriptor by incorporating discriminative information among color bands [4], which outperforms texture features, such as Gabor texture [7] and local binary pattern (LBP) [8]. There are also other global features for RSIR like simple statistics [9], GIST features [10], and Gray-Level Co-occurrence Matrix (GLCM) features [11].

Unlike global features, local features are generally captured from image patches centered at points of interest, and often achieve better performance than global features. SIFT is the most popular local descriptor, and has been used widely for various applications, including RSIR. Yang et al. released the first remote sensing benchmark dataset to the public and investigated the performance of local invariant features for RSIR [9]. The local features outperformed global features, such as simple statistics, color histograms, and texture features. Özkan et al. investigated the performance of state-of-the-art representation methods for geographical image retrieval [12]. Their extensive experiments indicate the advantages of local features for RSIR. However, local features like SIFT are of high dimension, and thus feature aggregation approaches, such as bag of visual words (BOVW) [13], vector of locally aggregated descriptors (VLAD) [14], and improved fisher kernel (IFK) [15] are often used to encode local features to generate more compact global features. Compared with VLAD and IFK, BOVW is not only an image representation widely used for RSIR [9,12], but also a framework that can combine with other features to extract more powerful feature representations [16,17]. Some other popular local features include histogram of oriented gradient (HOG) [18], and its variant descriptor pyramid histogram of oriented gradient (PHOG) [19].

Deep learning has been demonstrated to be capable of extracting more powerful feature representations compared to handcrafted features. The remote sensing community, and more specifically RSIR, has benefited from these deep learning approaches, since retrieval performance is greatly dependent on the effectiveness of feature representations as mentioned above. Zhou et al. proposed an unsupervised feature learning approach where SIFT and a sparse auto-encoder are combined to learn sparse features for RSIR [20]. In a recent work, Wang et al. proposed a novel graph-based learning method for effectively retrieving remote sensing images based on a three-layer framework [21]. The improvement of these two unsupervised feature learning methods, however, are limited since they are made based on shallow networks that cannot learn higher-level information.

In addition to unsupervised feature-learning-based methods mentioned above, convolutional neural networks (CNNs) are supervised ones that have been proved to be the most successful deep-learning approach based on their remarkable performance achieved on those benchmark datasets, such as ImageNet [22]. However, a large number of labeled images are needed to train effective CNNs from scratch, which is impossible for some domains (e.g., remote sensing) due to the lack of large-scale labeled datasets. In practice, transfer learning is often used to remedy the lack of labeled datasets by either treating the pre-trained CNNs as feature extractors, or fine-tuning the pre-trained CNNs on the target dataset. Napoletano presented an extensive evaluation of visual descriptors, including global, local, and CNN-based features [23]. The results demonstrate that features extracted by treating pre-trained CNNs as feature-extractors are able to achieve the best performance. Zhou et al. proposed a low dimensional convolutional neural network (LDCNN) based on convolutional layers and a three-layer perceptron, which can learn low-dimensional features from limited labelled images [24]. The Visual Geometry Group (VGG) networks [25], including three CNN models, i.e., VGGF, VGGM, and VGGS, have been investigated as the basic convolutional blocks of LDCNN, among which, VGGM performs the best on several benchmark datasets.

2.2. Multi-Label RSIR

The single-label RSIR methods mentioned above are effective in searching remote sensing images of interest from a large-scale archive, but the primitive classes (multiple labels) present in images are

ignored. This may result in a poor performance due to the semantic gap between low-level features and high-level concepts. Multi-label RSIR is different from single-label RSIR in terms of the number of labels included in images, as well as the process of feature extraction. In addition, for multi-label RSIR, a two-step coarse-to-fine retrieval can be performed based on the multiple labels in each image. More specifically, in the coarse retrieval step, the images in the archive that have at least one overlapped label with a query image will be returned to form the similar subset, and later in the fine retrieval step, the features extracted from the segmented image regions are used to perform exact retrieval of similar images from the subset. Figure 1 shows a basic comparison between single-label and multi-label RSIR.

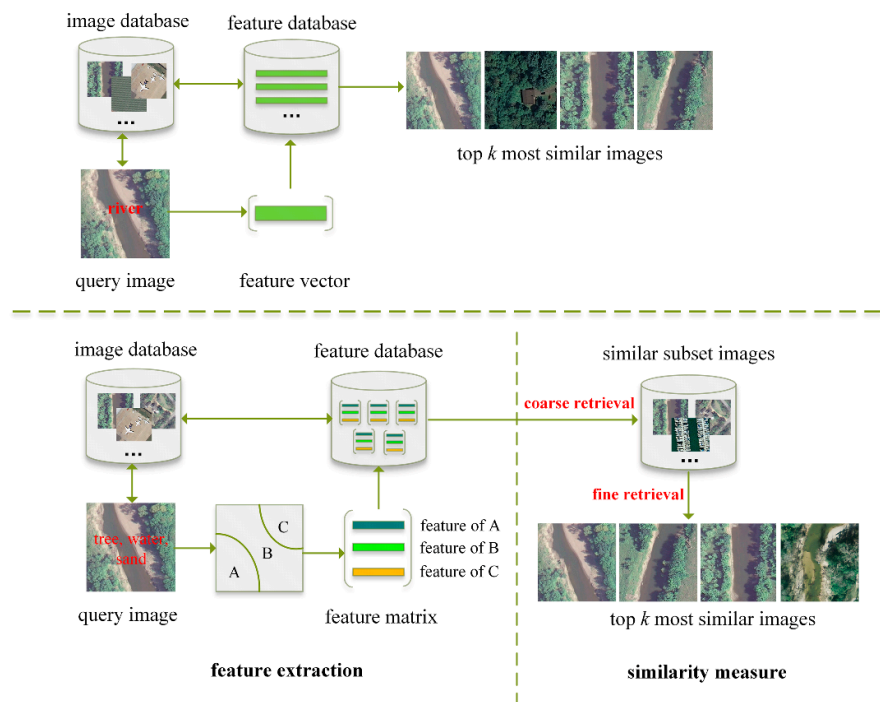


Figure 1. Comparison of single-label and multi-label RSIR.

To exploit the multiple labels and further improve RSIR performance, multi-label learning has shown promising and effective performance when it comes to addressing multi-label image retrieval problems in computer vision literature [26–28]. Nasierding et al. investigated multi-label classification methods for image annotation and retrieval to give a comparative study of these methods [26]. Li et al. proposed a novel multi-label image annotation method for image retrieval based on annotated keywords [27]. The results indicate that multi-labels can provide abundant descriptions for image content at the semantic level, thus improving precision and recall of image retrieval. Ranjan et al. introduced multi-label canonical correlation analysis to address cross-modal retrieval problem in the presence of multi-label annotations [28]. The proposed cross-model retrieval method achieves state-of-the-art retrieval performance.

Inspired by the success of multi-label learning methods in computer vision literature, the remote sensing community has raised interest in multi-label learning for RSIR problems [29–33]. Wang et al. proposed a remote sensing image retrieval scheme by using image scene semantic matching [29], and in the other work [30], image visual, object, and spatial relationship semantic features are combined to perform a two-stage coarse-to-fine retrieval of remote sensing images from multiple sensors. However, an object-based support vector machine (SVM) classifier is needed to produce classification maps of query images and images to be retrieved in the archive. In order to train an effective classifier, a reliable pixel-based training set is required, which is, however, not efficient for RSIR applications. Chaudhuri et al. presented a novel unsupervised graph-theoretic approach for region-based retrieval

of remote sensing images [31]. In the proposed approach, the images are modeled by an attributed relational graph, and then the graphs of the images in the archive are matched to that of the query image based on inexact graph matching. Dai et al. explored the use of multiple labels for hyperspectral image retrieval and presented a novel multi-label RSIR system combining spectral and spatial features [32]. Experimental results obtained using a benchmark archive of hyperspectral images show that the proposed method was successful for the adaptation of single-label classification for multi-label RSIR. In a recent work, Chaudhuri et al. proposed a multi-label RSIR method using a semi-supervised graph-theoretic method [33], which is an improvement of the region-based retrieval approach [31]. The proposed approach requires only a small number of pixel-wise labeled training images characterized by multiple labels to perform a coarse-to-fine retrieval process. This work provides not only a multi-label benchmark dataset but also baseline results for multi-label RSIR.

3. MLRSIR: A Pixel-Wise Dataset for Multi-Label RSIR

For single-label RSIR, a number of benchmark datasets are publicly available [34]. However, few works have been done to release datasets for multi-label RSIR in the remote sensing literature, which limits the development of novel approaches. Chaudhuri et al. released a multi-label RSIR archive [33], and each image in this archive is manually labeled with one or more labels based on visual inspection. This is the first open-source dataset for multi-label RSIR. However, it is an image-level dataset, which is sufficient for unsupervised or semi-supervised multi-label RSIR, but has limitations in supervised deep learning approaches, such as fully convolutional networks (FCN) [35]. More specifically, we only know the labels/primitive classes for the images but have no idea of the pixel-wise labels in each image.

As the initial step of the semi-supervised approach [33], an effective segmentation algorithm is required to obtain a number of semantically meaningful regions, since the retrieval performance is heavily dependent on the accuracy of segmentation results. During the subsequent steps, a small number of training images are randomly selected and pixel-wise labeled to predict the label of each region in an image. These steps can be combined and replaced by a FCN network, which has been proved to be effective for addressing semantic segmentation problem. Moreover, it is worth noting that pixel-wise labeling is also required in the semi-supervised multi-label RSIR approach. We therefore propose a new pixel-wise labeling dataset termed MLRSIR for multi-label RSIR that can be used for not only unsupervised and semi-supervised approaches but also supervised approaches like FCN.

3.1. Description of MLRSIR

To be consistent with the multi-label RSIR archive [33], the total number of distinct class labels associated for MLRSIR was also 17. The eCognition 9.0 (<http://www.ecognition.com/>) software was used to segment each image in the UC Merced archive [9] into a number of semantically meaningful regions, and then each region was assigned one of 17 pre-defined class labels.

MLRSIR had a total number of 21 broad categories with 100 images per class, which is the same as the UC Merced archive. The following 17 class labels, i.e., airplane, bare soil, buildings, cars, chaparral, court, dock, field, grass, mobile home, pavement, sand, sea, ship, tanks, trees, and water, were considered in this dataset. Figure 2 shows some images with corresponding pixel-wise labeling results, and the total number of images associated for each class label is shown in Table 1.

MLRSIR was a pixel-wise labeled dataset with each image containing multiple labels, therefore, it could also be used for other tasks, such as semantic segmentation (also called classification in remote sensing) and multi-label classification, i.e., predicting the classes contained in an image. MLRSIR is available at <https://sites.google.com/view/zhouwx/dataset>.

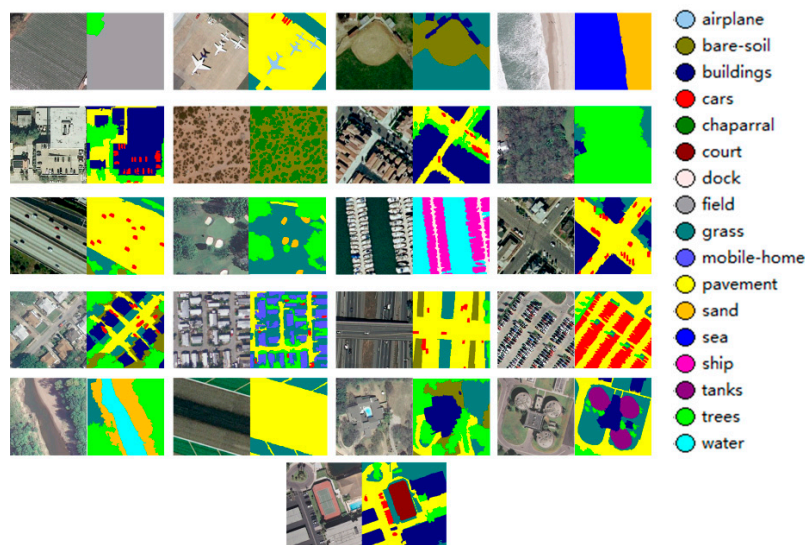


Figure 2. Example images and corresponding labeling results.

Table 1. The number of images present in the dataset for each class label.

Class Label	Number of Images
airplane	100
bare soil	754
buildings	713
cars	897
chaparral	116
court	105
dock	100
field	103
grass	977
mobile home	102
pavement	1331
sand	291
sea	101
ship	103
tanks	100
trees	1021
water	208

3.2. Multi-Label RSIR Based on Handcrafted and CNN Features

Multi-label RSIR was different from single-label RSIR in that for multi-label RSIR, the multi-label information was considered and the features are extracted from the segmented regions instead of the whole image. This section introduces the handcrafted and CNN features that were evaluated using the presented MLRSIR dataset.

3.2.1. Multi-Label RSIR Based on Handcrafted Features

To extract handcrafted features, we first determined the number of connected regions in each image according to its corresponding labeling results. We then extracted features from each of the segmented regions and combined these region-based features to form a feature matrix, as shown in Figure 1. In detail, each region was represented by a feature vector concatenating color, texture, and shape features. We refer the readers to Section 4.1 for more details on handcrafted feature extraction.

Two schemes were proposed to evaluate the retrieval performance of handcrafted features. In the first scheme, the multi-label RSIR was evaluated as single-label RSIR. More specifically, the similarity

between the query image and other images in the archive were obtained by calculating the distance between corresponding feature matrices as follows:

$$D(v_q, v_r) = \frac{1}{n} \sum_{q=1}^n \min(D(q, r)) \quad (1)$$

where v_q and v_r were the features of the query image and other images in the archive, respectively. $D(q, r)$ (D is a distance matrix) was the L_1 distance between the region q of the query image and region r of other images in the archive, and n was the number of regions in the query image. The first scheme is termed MLIR hereafter for conciseness.

In the second scheme, we performed a coarse-to-fine retrieval process. For the coarse retrieval step, a subset consisting of images which have at least one overlapped label with the query image was first obtained by comparing the label vectors (17-D vector) between the query image and other images in the archive. Then, in the later fine-retrieval step, we repeated the first scheme mentioned above on the subset to further improve retrieval results. The second scheme is termed MLIR-CF hereafter for conciseness.

3.2.2. Multi-Label RSIR Based on CNN Features

For multi-label RSIR based on CNN features, the pre-trained CNNs were fine-tuned on the MLRSIR dataset to learn domain-specific features. It is worth noting that the label of each image was a 17-D vector with the entries of 1 s and 0 s, where 1 indicated the image has this class, and 0 otherwise.

To evaluate CNN features for multi-label RSIR, we extracted the features from the fine-tuned fully-connected layers and proposed two schemes to investigate the performance. In the first scheme, the CNN-features-based multi-label RSIR was also evaluated as a single-label RSIR, which was the same as the first scheme in Section 3.2.1.

The second scheme relied on the label vectors to perform a coarse retrieval. Specifically, we first split the MLRSIR archive into two subsets, i.e., training and test sets, respectively, where the training set was used to fine-tune the pre-trained CNN, while the test set was used to perform coarse retrieval. Then we predicted the label vector of each image in the test archive by converting its corresponding label score (the output of the fine-tuned CNN) into binary values (0 and 1).

For binarization, a 17-D threshold vector was needed. Let $L = [l_{i,1}, l_{i,2}, l_{i,3}, \dots, l_{i,k}] (i = 1, 2, \dots, n)$ and $S = [s_{i,1}, s_{i,2}, s_{i,3}, \dots, s_{i,k}] (i = 1, 2, \dots, n)$ denote the label vectors and corresponding label scores of all the training images, respectively, where n and k were the number of training images and class labels, respectively. For class label k , the threshold t_k was determined by taking the average of the minimum label score with $l_{i,k} = 1$, and the maximum label score with $l_{i,k} = 0$, of all the training images. This process was repeated 17 times to obtain the 17-D threshold vector. Then the class label l_k of each test image was set to 1 if $l_k \geq t_k$, and 0 otherwise. Once the label vectors of all the test images were obtained, the hamming distance between the query image and other images in the archive was calculated by comparing their corresponding label vectors, as shown in Equation (2), where l_q and l_r were the label vector of the query image and other images in the archive, respectively, and L was the number of class labels. The second scheme is termed CNN-HM hereafter for conciseness.

$$Hamming = \frac{XOR(L_q, L_r)}{L} \quad (2)$$

4. Experiments and Results

In this section, we evaluate the single-label and multi-label RSIR methods on the proposed MLRSIR dataset.

4.1. Experimental Setup

In our experiments, simple statistics, color histogram, Gabor texture, HOG, PHOG, GIST, and LBP were used for single-label RSIR based on handcrafted features, as well as evaluated on the presented MLRSIR dataset. For the multi-label RSIR based on handcrafted features, i.e., MLIR, each region was described by concatenating color (histogram of each channel), texture (GLCM), and shape features (area, convex area, perimeter, extent, solidity, and eccentricity) to obtain a 110-D feature vector.

For the single-label and multi-label RSIR based on CNN features, we chose VGGM as the pre-trained CNN since it was able to achieve slightly better performance than the other two VGG networks, i.e., VGGF and VGGS, on the UC Merced archive. The VGGM network was fine-tuned with a single label and multiple labels, respectively. The convolutional architecture for fast feature embedding (Caffe) framework [36] was used for fine-tuning, and the parameters are shown in Table 2. In addition, the weights of the pre-trained VGGM were transferred to the network to be fine-tuned. To accelerate training and avoid overfitting, the weights of convolutional layers were fixed during fine-tuning. The weights of the first two fully-connected layers were used as initial weights, and the weights of the last fully-connected layer were initialized from a Gaussian distribution (with a mean of 0 and a standard deviation of 0.01). We randomly selected 80% of the images from each broad category of MLRSIR as the training set, and the remaining 20% of the images were used for evaluating retrieval performance.

Table 2. Fine-tuning parameters for single-label and multi-label CNN.

Parameters	Single-Label CNN	Multi-Label CNN
base learning rate	0.0001	0.0001
momentum	0.9	0.9
weight decay	0.0005	0.0005
max iterations	3000	4000

To be consistent with the recent work that presents a benchmark dataset to evaluate RSIR methods [34], we selected L_1 as the distance measure for the color histogram, and L_2 for other features. The average normalized modified retrieval rank (ANMRR), mean average precision (mAP), precision at k (P@k, k is the number of retrieved images), and precision-recall curve, were used to evaluate the retrieval performance. For ANMRR, the lower values indicated better performance, while for mAP and P@k, the larger the better. It is worth noting that each image was taken as a query image, and the query image itself was also regarded as a similar image in the following experiments.

To further evaluate the performance of the multi-label RSIR methods, three metrics, i.e., accuracy, precision, and recall, were computed. The equations are as follows:

$$Accuracy = \frac{1}{N} \sum_{i=1}^N \frac{|Q \cap R_i|}{|Q \cup R_i|} \quad (3)$$

$$Precision = \frac{1}{N} \sum_{i=1}^N \frac{|Q \cap R_i|}{|R_i|} \quad (4)$$

$$Recall = \frac{1}{N} \sum_{i=1}^N \frac{|Q \cap R_i|}{|Q|} \quad (5)$$

where N and L were the number of returned images and labels, respectively. Q and R_i were the label vector of the query image and the i th returned image, respectively.

4.2. Experimental Results

4.2.1. Results of Single-Label and Multi-Label RSIR

The single-label and multi-label RSIR based on handcrafted features were evaluated using the whole MLRSIR dataset, and the results are shown in Table 3. It can be observed that the multi-label RSIR method MLIR outperforms most of the handcrafted features except Gabor texture feature which achieves the best performance in terms of ANMRR value. However, we can see MLIR tends to achieve slightly better performance than Gabor texture in terms of P@k values when the number of returned images increases ($k \geq 1000$), indicating MLIR is more scalable than Gabor texture in a large-scale remote sensing archive. The results in Table 3 demonstrate the advantages of multi-label RSIR over single-label RSIR.

Table 3. Performance of single-label and multi-label RSIR based on handcrafted features. The bold values mean the best result for each performance metric.

Handcraft Features	ANMRR	mAP	P@5	P@10	P@50	P@100	P@1000
Simple Statistics	0.8580	0.1027	0.2843	0.1912	0.1138	0.0994	0.0646
Color Histogram	0.7460	0.1918	0.6489	0.5076	0.2694	0.1964	0.0680
Gabor Texture	0.7070	0.2232	0.6771	0.5508	0.3085	0.2286	0.0680
HOG	0.8600	0.1233	0.3415	0.2449	0.1322	0.1037	0.0564
PHOG	0.7990	0.1557	0.4976	0.3840	0.2105	0.1538	0.0581
GIST	0.7760	0.1859	0.5755	0.4540	0.2447	0.1756	0.0616
GLCM	0.7700	0.1545	0.4730	0.3744	0.2220	0.1727	0.0683
LBP	0.7710	0.1648	0.6158	0.4845	0.2589	0.1810	0.0580
MLIR	0.7460	0.2029	0.5364	0.4267	0.2558	0.1985	0.0703

Table 4 shows the performance of single-label and multi-label RSIR based on CNN features. These results were obtained using the test set, i.e., 20% of the MLRSIR dataset, as mentioned in Section 3.2.2. We extracted features from the first two fully-connected layers and obtained four features, i.e., CNN-Fc6, CNN-Fc6ReLU, CNN-Fc7, and CNN-Fc7ReLU. The results indicated that the fine-tuned CNN features outperformed the pre-trained CNN features, and that the multi-label RSIR performed slightly better than single-label RSIR for these four features except CNN-Fc7. The ANMRR values of CNN-Fc7 were 0.3350 and 0.3440 for single-label and multi-label RSIR, respectively. It can also be observed that the activation function ReLU affected the performance of features extracted from the fully-connected layers for both single-label and multi-label RSIR. In addition, CNN-HM achieved the worst performance for all the evaluated performance metrics. This was because of the fact that CNN-HM is essentially a coarse retrieval that only relied on the labels of the images. CNN-HM could be used as the first-stage retrieval to filter out those images that did not contain the specific classes as the query image.

Figure 3 shows the precision-recall curves for single-label and multi-label RSIR based on CNN features. The performance is consistent with the results in Table 4.

We selected the best performing features for the pre-trained CNN features, the single-label RSIR features and the multi-label RSIR features, respectively, and plotted the ANMRR histogram for each broad class in MLRSIR, as shown in Figure 4. We can see multi-label RSIR, i.e., CNN-Fc7ReLU (ML) in Table 4, achieved the best performance for most of the broad classes except intersection and parking lot. For an image class like intersection, multi-label RSIR even achieved the worst performance. A possible explanation is that the image of intersection usually contains more primitive classes, including pavement, cars, trees, grass, buildings, and bare soil. This made it difficult to accurately represent the images since the features were extracted from the regions, and we did not consider the spatial relationship between different regions.

Table 4. Performance of single-label and multi-label RSIR based on CNN features. “SL” and “ML” mean the CNNs are fine-tuned with single and multiple labels, respectively. “ReLU” means the feature is extracted with the use of activation function. The bold values mean the best result for each performance metric.

CNN Features	ANMRR	mAP	P@5	P@10	P@50	P@100
CNN-Fc6	0.3740	0.5760	0.7895	0.6624	0.2902	0.1758
CNN-Fc6ReLU	0.4050	0.5456	0.7629	0.6352	0.2781	0.1681
CNN-Fc7	0.3830	0.5619	0.7643	0.6417	0.2880	0.1714
CNN-Fc7ReLU	0.3740	0.5693	0.7814	0.6543	0.2906	0.1735
CNN-Fc6(SL)	0.3640	0.5862	0.7905	0.6714	0.2960	0.1780
CNN-Fc6ReLU(SL)	0.3680	0.5829	0.7900	0.6710	0.2936	0.1748
CNN-Fc7(SL)	0.3350	0.6123	0.8005	0.6979	0.3064	0.1794
CNN-Fc7ReLU(SL)	0.3180	0.6277	0.8233	0.7076	0.3113	0.1808
CNN-Fc6(ML)	0.3620	0.5870	0.7943	0.6767	0.2949	0.1773
CNN-Fc6ReLU(ML)	0.3700	0.5824	0.7810	0.6636	0.2899	0.1715
CNN-Fc7(ML)	0.3410	0.6074	0.7924	0.6879	0.3005	0.1745
CNN-Fc7ReLU(ML)	0.3220	0.6273	0.8076	0.7100	0.3080	0.1777
CNN-HM	0.4270	0.5188	0.6052	0.5676	0.2713	0.1617

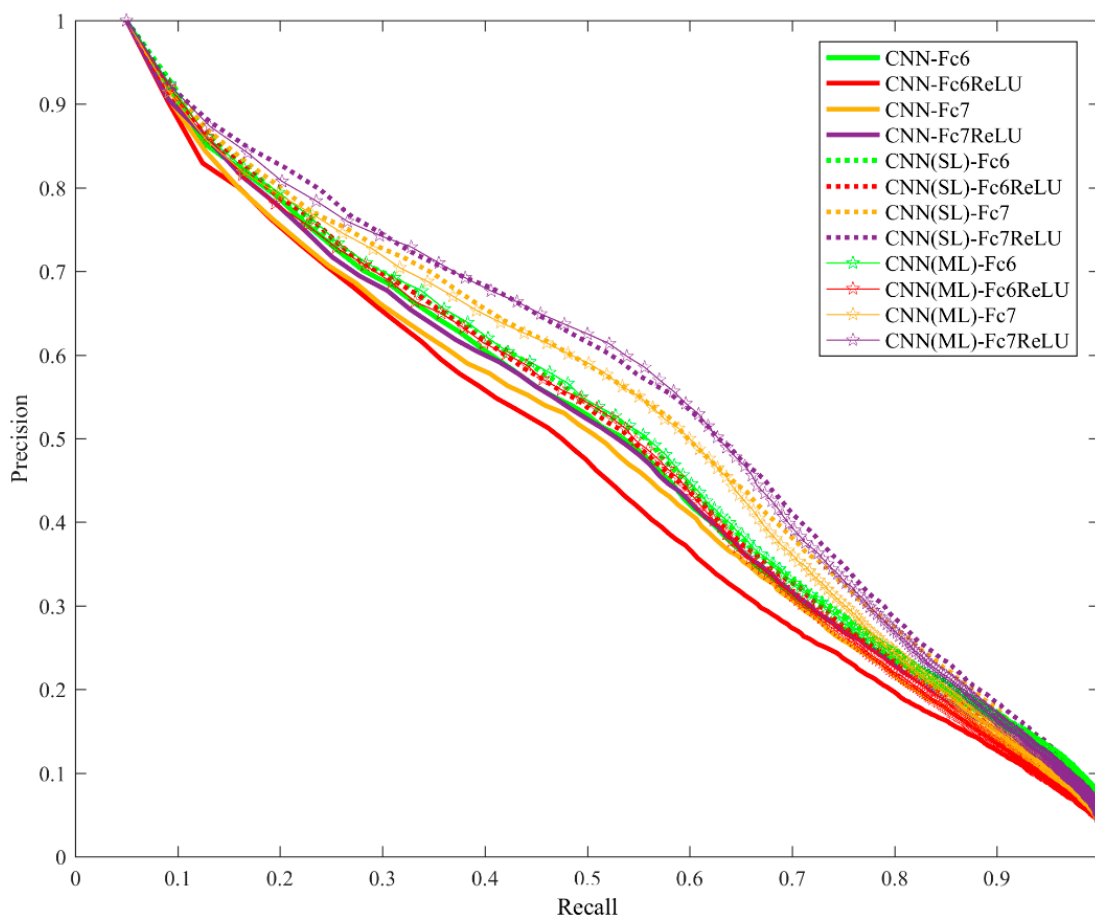


Figure 3. The precision-recall curves for single-label and multi-label RSIR.

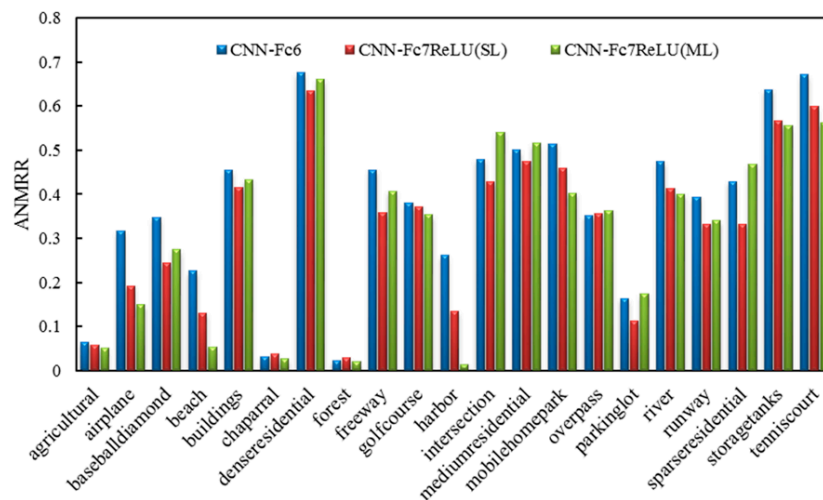


Figure 4. The results of CNN features for each class in MLRSIR.

4.2.2. Comparisons of the Multi-Label RSIR Methods

We compared our multi-label RSIR method (MLIR-CF) with several state-of-the-art methods, including KNN, ARGMM, and MLIRM, on the presented MLRSIR dataset. The results are shown in Table 5. It can be seen that MLIR-CF outperformed KNN, but performed worse than the other two methods. This is because the graph matching strategy based on an attributed relational graph (ARG) was used for the similarity measure in both ARGMM and MLIRM.

Table 5. Comparisons of multi-label RSIR methods. The bold values mean the best result for each performance metric.

Features	Accuracy	Precision	Recall
KNN [33]	0.5218	0.6397	0.6102
ARGMM [33]	0.6356	0.7234	0.6987
MLIRM [33]	0.7429	0.8568	0.8025
MLIR-CF	0.6188	0.6813	0.8177

5. Conclusions

In this paper, we proposed a benchmark dataset named MLRSIR for multi-label RSIR. We expect MLRSIR to help advance the development of RSIR approaches, particularly supervised-learning-based methods. We also compared the performance of single-label and multi-label RSIR on MLRSIR based on handcrafted and CNN features. MLRSIR is collected for RSIR and particularly multi-label RSIR, but it can also be used for other problems such as semantic segmentation.

Author Contributions: The research idea and design were conceived by K.Y. and W.Z. The experiments were performed by K.Y. and W.Z. The manuscript was written by K.Y. Z.S. helped revise the manuscript.

Acknowledgments: This work was supported by the National key research and development plan on a Strategic International Scientific and Technological Innovation Cooperation Special Project (2016YFE0202300), Wuhan Chen Guang Project (2016070204010114), Guangzhou Science and Technology Project (201604020070); Special Task of Technical Innovation in Hubei Province (2016AAA018), and the Natural Science Foundation of China (61671332, 41771452 and 41771454). The authors would like to thank the anonymous reviewers for their hard work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bosilj, P.; Aptoula, E.; Lefèvre, S.; Kijak, E. Retrieval of Remote Sensing Images with Pattern Spectra Descriptors. *ISPRS Int. J. Geo-Inf.* **2016**, *5*, 228. [[CrossRef](#)]
2. Aptoula, E. Remote sensing image retrieval with global morphological texture descriptors. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 3023–3034. [[CrossRef](#)]
3. Bouteldja, S.; Kourgli, A. Multiscale texture features for the retrieval of high resolution satellite images. In Proceedings of the 2015 International Conference on Systems, Signals and Image Processing (IWSSIP), London, UK, 10–12 September 2015; pp. 170–173.
4. Shao, Z.; Zhou, W.; Zhang, L.; Hou, J. Improved color texture descriptors for remote sensing image retrieval. *J. Appl. Remote Sens.* **2014**, *8*, 83584. [[CrossRef](#)]
5. Scott, G.J.; Klaric, M.N.; Davis, C.H.; Shyu, C.R. Entropy-balanced bitmap tree for shape-based object retrieval from large-scale satellite imagery databases. *IEEE Trans. Geosci. Remote Sens.* **2011**, *49*, 1603–1616. [[CrossRef](#)]
6. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [[CrossRef](#)]
7. Liu, C.; Wechsler, H. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE Trans. Image Process.* **2002**, *11*, 467–476. [[PubMed](#)]
8. Ojala, T.; Pietikainen, M.; Maenpaa, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 971–987. [[CrossRef](#)]
9. Yang, Y.; Newsam, S. Geographic image retrieval using local invariant features. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 818–832. [[CrossRef](#)]
10. Oliva, A.; Torralba, A. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vis.* **2001**, *42*, 145–175. [[CrossRef](#)]
11. Howarth, P.; Rüger, S. Evaluation of texture features for content-based image retrieval. In Proceedings of the International Conference on Image and Video Retrieval, Dublin, Ireland, 21–23 July 2004; Springer: Berlin, Germany, 2004; pp. 326–334.
12. Özkan, S.; Ateş, T.; Tola, E.; Soysal, M.; Esen, E. Performance analysis of state-of-the-art representation methods for geographical image retrieval and categorization. *IEEE Geosci. Remote Sens. Lett.* **2014**, *11*, 1996–2000. [[CrossRef](#)]
13. Sivic, J.; Zisserman, A. Video Google: A text retrieval approach to object matching in videos. In Proceedings of the Ninth IEEE International Conference on Computer Vision, Nice, France, 13–16 October 2003; Volume 2, pp. 1470–1477.
14. Jégou, H.; Douze, M.; Schmid, C.; Pérez, P. Aggregating local descriptors into a compact image representation. In Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010; pp. 3304–3311.
15. Perronnin, F.; Sánchez, J.; Mensink, T. Improving the Fisher Kernel for Large-Scale Image Classification. In Proceedings of the European Conference on Computer Vision, Crete, Greece, 5–11 September 2010; pp. 143–156.
16. Yang, J.; Liu, J.; Dai, Q. An improved Bag-of-Words framework for remote sensing image retrieval in large-scale image databases. *Int. J. Digit. Earth* **2015**, *8*, 273–292. [[CrossRef](#)]
17. Aptoula, E. Bag of morphological words for content-based geographical retrieval. In Proceedings of the 2014 12th International Workshop on Content-Based Multimedia Indexing (CBMI), Klagenfurt, Austria, 18–20 June 2014; pp. 1–5.
18. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893.
19. Bosch, A.; Zisserman, A.; Munoz, X. Representing shape with a spatial pyramid kernel. In Proceedings of the 6th ACM International Conference on Image and Video Retrieval, Amsterdam, The Netherlands, 9–11 July 2007; ACM: New York, NY, USA, 2007; pp. 401–408.
20. Zhou, W.; Shao, Z.; Diao, C.; Cheng, Q. High-resolution remote-sensing imagery retrieval using sparse features by auto-encoder. *Remote Sens. Lett.* **2015**, *6*, 775–783. [[CrossRef](#)]

21. Wang, Y.; Zhang, L.; Tong, X.; Zhang, L.; Zhang, Z.; Liu, H.; Xing, X.; Mathiopoulos, P.T. A three-layered graph-based learning approach for remote sensing image retrieval. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 6020–6034. [[CrossRef](#)]
22. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Li, F.-F. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 2–9.
23. Napoletano, P. Visual descriptors for content-based retrieval of remote-sensing images. *Int. J. Remote Sens.* **2018**, *39*, 1343–1376. [[CrossRef](#)]
24. Zhou, W.; Newsam, S.; Li, C.; Shao, Z. Learning Low Dimensional Convolutional Neural Networks for High-Resolution Remote Sensing Image Retrieval. *Remote Sens.* **2017**, *9*, 489. [[CrossRef](#)]
25. Chatfield, K.; Chatfield, K.; Simonyan, K.; Vedaldi, A.; Zisserman, A. Return of the Devil in the Details: Delving Deep into Convolutional Nets. In Proceedings of the British Machine Vision Conference, Nottingham, UK, 1–5 September 2014; pp. 1–11.
26. Nasierding, G.; Kouzani, A.Z. Empirical study of multi-label classification methods for image annotation and retrieval. In Proceedings of the 2010 International Conference on Digital Image Computing: Techniques and Applications (DICTA), Sydney, NSW, Australia, 1–3 December 2010; pp. 617–622.
27. Li, R.; Zhang, Y.; Lu, Z.; Lu, J.; Tian, Y. Technique of image retrieval based on multi-label image annotation. In Proceedings of the 2010 Second International Conference on Multimedia and Information Technology (MMIT), Kaifeng, China, 24–25 April 2010; pp. 10–13.
28. Ranjan, V.; Rasiwasia, N.; Jawahar, C.V. Multi-label cross-modal retrieval. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4094–4102.
29. Wang, M.; Song, T. Remote sensing image retrieval by scene semantic matching. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 2874–2886. [[CrossRef](#)]
30. Wang, M.; Wan, Q.M.; Gu, L.B.; Song, T.Y. Remote-sensing image retrieval by combining image visual and semantic features. *Int. J. Remote Sens.* **2013**, *34*, 4200–4223. [[CrossRef](#)]
31. Chaudhuri, B.; Demir, B.; Bruzzone, L.; Chaudhuri, S. Region-based retrieval of remote sensing images using an unsupervised graph-theoretic approach. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 987–991. [[CrossRef](#)]
32. Dai, O.E.; Demir, B.; Sankur, B.; Bruzzone, L. A novel system for content based retrieval of multi-label remote sensing images. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, 23–28 July 2017; pp. 1744–1747.
33. Chaudhuri, B.; Demir, B.; Chaudhuri, S.; Bruzzone, L. Multilabel Remote Sensing Image Retrieval Using a Semisupervised Graph-Theoretic Method. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 1144–1158. [[CrossRef](#)]
34. Zhou, W.; Newsam, S.; Li, C.; Shao, Z. Patternet: A benchmark dataset for performance evaluation of remote sensing image retrieval. *arXiv*, 2017. [[CrossRef](#)]
35. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
36. Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; Darrell, T. Caffe: Convolutional Architecture for Fast Feature Embedding. In Proceedings of the ACM International Conference on Multimedia, Orlando, FL, USA, 3–7 November 2014; ACM, 2014; pp. 675–678.

