A New Remote Sensing Image Retrieval Method Based on CNN and YOLO

Junwei Xin[1*], *Famao Ye*[1], *Yuanping Xia*[2], *Yan Luo*[1], *Xiaoyong Chen*[1]

[1] *State Key Laboratory of Nuclear Resources and Environment, East China University of Technology, China*
[2] *School of Surveying and Mapping Engineering, East China University of Technology, China*
xinjunwei@ecut.edu.cn, yefamao@ecut.edu.cn, 200860011@ecut.edu.cn, yanluo@ecut.edu.cn, 200960002@ecut.edu.cn

Abstract

Retrieving remote sensing images plays a key role in RS fields, which activates researchers to design a highly effective extraction method of image high-level features. However, despite the advanced features of the image can be extracted by deep-learning method, features fail to cover the overall information due to its rich and complex background. Extract key regions from RS images, recede background interference and retrieval accuracy needs to be solved. Combined YOLOv5 target recognition algorithm with deep-learning method, this paper proposes a novel retrieval method based on target critical region detection. Firstly, the key retrieval regions have been identified. YOLOv5 target recognition algorithm has been used to identify the key regions of the image and served as the retrieval regions. Secondly, the retrieval characteristics are determined. Combining with the CNN model ResNet50, the retrieval features are extracted from the retrieval regions acquired in the previous step, in addition, PCA method has been used to reduce the dimension of the retrieval features. Finally, using weighted distance based on class probability to measure the similarity between a query and retrieve images. Experimental results show that the proposed method can extract better image retrieval features and improve the retrieval performance of RS image.

Keywords: RS image retrieval, CNN, PCA, Weighted distance, Key region

1 Introduction

With the accelerated development of aerospace technology and sensor technology, it is more and more convenient to acquire image data with diversification. RS has been widely used in environmental monitoring, agricultural monitoring, urban planning, disaster management and many other fields [1]. However, differ to its acquisition speed, the existing processing technology of RS data is fail to meet the requirements, resulting into a great waste of RS data. How to obtain effective information from the massive RS data, efficiently and quickly retrieve the images of interest is one of the important tasks in the application of RS [2]. The main purpose of RS image retrieval is to find the images that fit the requirements from the massive image data set and sort descending in similarity. However, RS data has the characteristics of massive, complexity and diversity, which proposes a higher request in the image retrieval. Strategy for content-based RS image retrieval (CBRSIR) is a study hotspot. The quality of RS image retrieval mainly be decided by the ability of image features expression [3]. Extraction of useful retrieval features is a thorny problem in RS image retrieval. With the increasing demand for retrieval quality, the feature extraction of RS image retrieval has been put forward higher requirements [4-5].

Relying on the way of features extraction, features can generally be divided into two categories [6]. The first category is the traditional hand-crafted descriptors that include global hand-crafted features and local hand-crafted descriptors. In the early stage of CBRSIR, global hand-crafted features are frequently used as spectral [7] and shape features [8]. This method doesn’t rely on manual annotation and extracts features directly. RS image retrieval can be realized through similarity matching between features. Zhang et al. [9] used gray level co-occurrence matrix to extract spectral and texture features for Hyperspectral remote sensing image retrieval. Local features can be encoded into compact global image representation by feature coding technology, and then local manual descriptors has been formed. Bag of Words (BoW) has shown outstanding performance in image retrieval [10]. Image local features were encoded into a quantized histogram feature model constructed by K-means clustering method. Yang et al. [11] features extracted for high resolution RS images were retrieved using Bag of Words (BoVW) coding method. In [12], The VLAD and BoVW were compared in satellite image retrieval. These methods are mainly using to extract underlying features for remote sensing images, which cann’t fully reflect accurate image semantic information, resulting in semantic gaps and a lack of deeper understanding of image features. Therefore, it is difficult to extract the required features information from massive data effectively.

Compared to the first category above, as the method based on deep-learning has made outstanding achievements in the field of target recognition, the deep-learning methods have cause the concern of the RS research community [13]. With good distinguishing ability and robustness, the image retrieval algorithm based on deep-learning can extract more features. Similar to the

*Corresponding Author: Junwei Xin; E-mail: xinjunwei@ecut.edu.cn

DOI: 10.53106/160792642023032402002

human brain, deep-learning builds a machine learning architecture model, which contains more than one hidden layer [13]. From that, we can automatically learn image Intrinsically abstract mathematical relationship, and extracts the image features from the bottom to the more discriminating higher levels of the image step by step from the input data. Finally, it can provide an effective framework for automatic extraction of target characteristics, and it has more research value for mining geographic information data in RS image.

Under the background of RS image Big Data, the deep-learning model can reveal the information in massive RS data and obtain more representative image features through training of large-scale RS image data. It is a better method to extract features from RS images with a broad prospect for CBRSIR. CNN is one of the models of deep-learning and has made great achievements in object retrieval. Different from the model based on manual features, the model based on CNN is an end-to-end model, which trained by a large amount of data can automatically learn the rich information [14]. Therefore, the current image retrieval model is more based on CNN. [15-16] proposed some CNN models to extract features for CBRSIR. In [17], a new Network structure LDCNN based on multi-layer perceptron, which is essentially the combination of traditional CNN Network and NIN Network, adopts to RS image retrieval. Cui et al. [18] proposed a multi-scale RS image retrieval method that using deep-learning and complex spatial relationship features. Li et al. [19] proposed a CNN adversarial network to retrieve synthetic aperture radar image. In [20], a deep hash CNN model was proposed to effectively retrieve cross source RS images. In [21], a new method for Large-scale image retrieval has been proposed by using Symmetry, FAST Scores, shap-based filtering and space mapping combined with CNN. Desai et al. [22] present an efficient and fast deep learning framework for image retrieval which integrates both CNN and support vector machine (SVM).

RS images has many kinds of features and rich backgrounds, which will affect the retrieval accuracy. Lacking of understanding of RS image content, previous studies which focused on global feature extraction and ignore the background information, were always dividing inaccurately the region of interest and extracting ineffective content for retrieval finally. We firstly use the target recognition method to identify the key regions of RS images, and then we use CNN to extract the retrieval features of the contents of the key regions to reduce the interference of RS image background.

2 Methodology

On the principle of imitating human vision system, the thought of deep-learning is to stack multiple layers, and the output of the upper layer serves as the input of the lower layer. The multi-layer networks can be trained in this way, in order to realize the hierarchical expression of input information, with extracting features from the bottom, to the top of target shape or partially, and more to the higher level of the target and behavior of the target [14]. In order to extract more reliable retrieval features, we propose a new retrieval method used YOLOv5 to determine the retrieval region. Then image features can be extracted from the retrieval region for RS image retrieval.

Firstly, we could extract the key regions of an image by using a detection model YOLOv5, and then the key regions detected could be served as the search area for image retrieval. Secondly, by the classification ability of convolutional neural network we can extract the features in the retrieval area as the basic retrieval features. However, due to the vector dimension of the extracted basic retrieval features is too high, we use PCA method to reduce the dimension of features, for that it can reduce the retrieval time and avoid the impact of redundant information between feature vectors. At last, we use weighted distance based on class probability as the similarity measure criterion to improve the retrieval accuracy.

2.1 Shadow Index

Due to the interference of RS image background on retrieval characteristics, this paper determines the key regions of image through the RS image target detection method based on YOLOv5. The most representative ground objects are detected by the YOLO v5 algorithm, and then the key regions of retrieval can be determined through these candidate features.

2.1.1 YOLOv5 Model

The YOLO algorithm creatively solves the target detection as a regression problem and obtains the detection frame boundary and category probability directly from the feature extraction layer. Different from RPN network or sliding window method, YOLO algorithm combines candidate region and detection stage into one. YOLO algorithm makes prediction based on the whole image and gives all detection results at one time. YOLO algorithm redefines the target detection task as a single regression prediction problem, which can directly obtain boundary box coordinates and category information from image pixels. Following the one-stage target detection algorithm route, the combined convolution layer structure is firstly used to extract multi-scale features. Then features can be fused in the full connection layer, and then the image feature tensor would be transferred to the prediction layer. Prediction module is used to process the result of grid prediction, and then the image features are recognized and classified. Finally, the boundary box of targets is generated and we can obtain category information for prediction.

According to the different network depth and dimension, YOLOv5 model is mainly divided into four models: s, m, l and x. Each model is different only in the width and depth of the network, and all models are composed of four parts: Input, Backbone, Neck, and Head. At the Input part, Mosaic data enhancement method is adopted to increase the number of small target samples by randomly scaling, clipping, stitching, and arranging the input images, so as to improve the detection accuracy of the model. The Backbone adopts Focus sub-sampling, CSP structure improved and SPP pooled pyramid structure to extract image feature information. The Neck mainly adopts the FPN + PAN feature pyramid structure to realize the transmission of feature information of targets of different sizes and solve the problem of multi-scale. The Head uses three loss functions to calculate classification, positioning and confidence loss respectively, and improves the accuracy of network prediction through NMS.

YOLOv5x model has more convolution kernels and can output deeper feature tensors, which has better detection effect. Under the test of the same data set, the average accuracy of detection results of YOLOv5x network is the best. The structure of YOLOv5x network is shown in Figure 1. In the first CSP1, the structure of YOLOv5x network includes 4 residual components, therefore is CSP1_4. In the second and third CSP1, 12 residual components are used, therefore is CSP1_12. In the Neck module, the CSP2 structure is also composed in the same way. After the first CSP2 structure, the YOLOv5x component structure includes two groups of four convolution operations, after eight convolution operations, so it is CSP2_4, YOLOv5x network structure diagram has been marked in detail, as shown in Figure 1. With the structural networks continuous deepening, the ability of feature extraction is strengthened, and fusion is also enhanced.

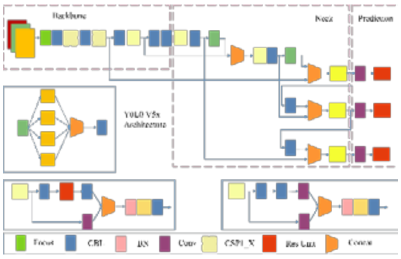

Figure 1. YOLOv5x network structure diagram

2.1.2 Determination of Object Location and Category by Using the YOLOv5

In the field of target detection, the prior frame can accurately reflect the size of the real target frame, so the determination of the prior frame is the key problem. Clustering algorithm is usually used to design prior frames. Clustering algorithm is a common unsupervised algorithm in data analysis. The clustering center generalizes the characteristics of a certain kind of data to a certain extent. K-means algorithm is a classical clustering algorithm, which judges the similarity between targets according to the distance. The closer the distance is, the greater the similarity is. In sense of RS images, it can be understood that the closer the distance is, the greater the degree overlap between the detection frame and the real target frame is.

For obtaining the location of the target objects in the image, K-means clustering method has been used to determine the prior frame firstly. Then after fine-tuning and refining the parameters (t_x, t_y, t_w, t_h) which are output of YOLOv5x network training, we can obtain the ground object positioning and the prediction frame, and the regression formula (1), (2) for predicting the center point and side length of the frame is as follows:

$$\begin{cases} b_x = \sigma(t_x) + c_x \\ b_y = \sigma(t_y) + c_y \end{cases}. \tag{1}$$

$$\begin{cases} b_w = p_w^{e^{tw}} \\ b_h = p_h^{e^{tw}} \end{cases}. \tag{2}$$

In the formula, b_x and b_y are the locations information of the center of predicted frame. In the feature graph, the upper-left coordinates of the cell network are c_x and c_y. And the b_w and b_h are the width and height attributes of the predicted border respectively, while p_w and p_h are the width and height attributes of the prior border respectively. The schematic diagram of border prediction is shown as Figure 2.

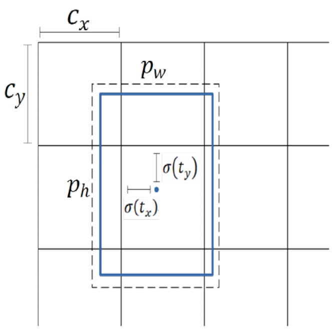

Figure 2. Border prediction schematic

Secondly, for obtaining the classification of the object, YOLOv5 algorithm is used to predict the confidence of the frame objects and determine the classification information of the objects. And then, we use the logistic activation function model to predict the border target category, and describe the classification problem with probability. To avoid overlapping of multiple labels in the targets, any category label greater than the set threshold can be tagged to the target in the prediction border.

$$S(x) = \frac{1}{1+e^{-x}}. \tag{3}$$

It can be seen from formula (3) that the mathematical model generated by this function is an s-shaped curve with a value range of (0, 1).

Finally, in order to verify the dependability of the classification task, we select the highest confidence score of the target class, and then use the predicted border of that to evaluate the target category. The fractional deviation between the predicted category output and the actual trained category is calculated using the binary cross-entropy loss function.

$$loss = -\frac{1}{n}\sum_{x}[yln a + (1-y)ln(1-a)]. \tag{4}$$

Where n denotes the total number and x represents the individual sample. While y expresses actual sample value, a is the predicted output value.

2.1.3 Key Regions Determination

There are many ground objects in the RS image, and the detection results of size and category of each ground object cannot be the same when using YOLO model. Taking Figure 3 as an example, which shows the practical detection case. Therefore, it is necessary to determine key region based on the detection results.

According to the image category information of a RS image M, we could obtain series classified object set G with its classification probability P as well as the center point location, and we could also get the height H and width W attributes, and the prediction box score S.

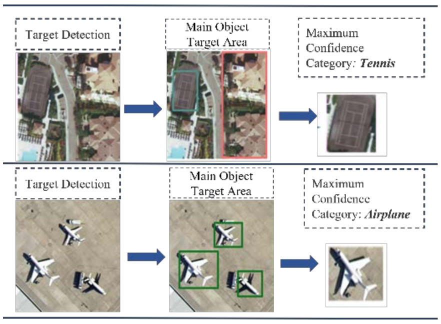

Figure 3. Determination of target area of main object

We would obtain many key regions of different sizes during the de-background of images, which would affect the retrieval accuracy of the features. In order to ensure that more main object targets would be included in the key region, the noise reduction threshold τ ($\tau \in [0, 1]$) is set to conduct multiscale subdivision and correction.

We would replace the key region's center point with the center point of the representative object. To make sure that the magnitude of key regions would be of the same, the width and height of the key region could be calculated by equation (5):

$$\begin{cases} H_N = H_O \times \tau \\ W_N = W_O \times \tau \end{cases}. \tag{5}$$

where H_O is the height, and W_O is the width, the value range of reduction coefficient τ is 0 to 1. When no representative object is detected, we used the image's center point as the key region's center point. In this way, we could obtain every image's key region.

2.2 Feature Extraction and Feature Dimension Reduction

Retrieval feature extraction is the key step in CBRSIR. Since YOLOv5 algorithm outputs convolution feature maps in three scales for detecting image targets. However, it is not suitable for retrieving images

ResNet is the latest network structure proposed by MSRA in 2015. Its biggest characteristic lies in the introduction of residual learning in the network. During the training of CNN model, as the network deepens, the gradient decreases continuously during forward transmission until it disappears, so that the parameters of the front layer cannot be updated, which affects the convergence of the network or makes the network fall into a poor local optimal solution. In ResNet, residual learning is introduced, that is, skip connection is added between some layers, so that features can be directly connected to deeper layers and gradients can be directly transferred from deeper layers to shallow layers, avoiding the problem of gradient disappearance. As a result, the performance of the network will not be affected with the deepening of the depth, and deeper high-level features can be extracted, which has a better advantage of image feature extraction. Therefore, we select a ResNet [23-24] model to extract retrieval features for RSIR after detecting the key region of an image.

However, because the dimension of feature vectors is too high, the amount of retrieval calculation will be increased. and the redundant information between them will also affect the retrieval accuracy. For reducing the influence of the redundant information of feature vectors and reduce the dimension, PCA method should also be used to reduce the dimension of feature vectors in key regions.

Firstly, we use ResNet50 residual network to learn the features of key regions, and extracted pool5 full-connection layer feature vector through the established model to represent the features of key regions. Secondly, we reduce the extracted features in dimension. Through PCA method, the dimension of selected features could be reduced to K.

2.3 Class-based Weighted Distance Method

Similarity measure is used to measure the similarity between two images, and the measure method has a great influence on the retrieval results. In the classification of RS image objects, the higher the probability of the image objects in a certain category, the higher the similarity degree in this category, and the smaller the distance between images. Therefore, the probability of RS image category is inversely proportional to the similarity measure distance.

Based on the powerful classification ability of CNN, the correlation between RS images of the same category is strengthened. For improving the retrieval accuracy of RS images, this paper uses the Class-based weighted distance method as the similarity measurement criterion to accurately characterize and quantify the similarity between images. The Class-based weighted distance method proposed in [25] believes that the greater the probability that the retrieved image belongs to the query image, the greater the weight value assigned.

For a query image q, image classification probability p^q can be obtained by ResNet network model. When performing image retrieval task, distance weight w_r that query image q and retrieval image r can be obtained by formula (6).

$$W_r = 1 - p_q^{cr}. \tag{6}$$

Similarly, distance weight w_q of retrieval image r and query image q can be obtained, and formula (7) is as follows:

$$W_q = 1 - p_r^{cq}. \tag{7}$$

Where, cr is the category of image r to be retrieved, and cq is the category of image q to be queried.

According to formula (7) and similarity measure distance, the weighted distance of the two image can be calculated by using equation (8).

$$WD(q,r) = W_r \times W_q \times d_{(q,r)}. \quad \textbf{(8)}$$

In the formula (8), $d_{(q,r)}$ is Euclidean distance of the two images. According to the above formula, retrieved images would be sorted and then we obtain the retrieval result finally.

2.4 Retrieval Algorithm Process

The RS image retrieval methods proposed here mainly include RS image retrieval region detection method by YOLOv5, retrieval region feature extraction method, PCA method and class-based weighted distance method.

The algorithm flow is divided into two sections: the Offline section and the Online section. The Offline section is mainly to complete the determination of the retrieval region, feature extraction and dimensionality reduction processing. The Online section is mainly to complete the calculation of similarity measurement distance and weighted sorting, and finally output the results of RS image retrieval. Table 1 shows the algorithm flows.

Table 1. Retrieve algorithm flow

RS image retrieval method based on YOLOv5
Offline Section: Input: Datasets of the retrieve images Step1: Fine-tune the YOLOv5 model with some labeled images Step 2: Use the fine-tuned YOLOv5 model to detect each image on the retrieval dataset Step 3: Determine the key region of each image on the retrieval dataset and extract the content on the key region Step 4: Use the fine-tuned ResNet50 model which was fine-tuned with some content extracted by the YOLOv5 model to extract the CNN features for every image on the retrieved dataset Output: A dataset with feature vectors and class labels
Online Section: Input: A query image is fed into the fine-tuned YOLOv5 model to obtain a set of ground objects O Step 1: Repeat offline steps 2, 3 and 4 to extract key area features of the query image Step 2: Reducing the dimension of the features by PCA method and estimating the class probability Step 3: The weighted distances based on class probability between the query image and the retrieved images are computed
Output: Sort the retrieved images according to the weighted distances and get the retrieval results

3 Experiment

3.1 Experimental Settings

In this paper, we selected the UC-Merced dataset [32] for experimental analysis, which includes 21 land-cover classes (The specific land-use categories are shown in Table 4), each class with 100 images. The spatial resolution is 0.3m and the image resolution is 256×256. The dataset contains rich ground object objects, the internal distance of the ground object category is large, with small space. And the ground object category of residential areas with different densities is difficult to distinguish, which is representative for the study of RS image retrieval algorithm.

To establish YOLOv5 model, we selected fifty images from each class randomly as training set, and the other fifty images are used for evaluating the retrieval performance. Parameters in the training were set as follow Table 2.

Table 2. Parameter settings in the training

Method	Parameters	Value
YOLOv5	Batch size	16
	Loss coefficient of GIOU	0.0306
	Loss coefficient of classification	0.211
	Threshold of label and anchor box	0.2
ResNet	Learning rate	0.001
	Momentum	0.9
	Weight decay	0.0005

To measure the performance of image retrieval system, it usually depends on the pre-established evaluation index rather than the way of manual participation. For evaluating the retrieval performance objectively and impartially, we almost use a combination of multiple evaluation mechanisms. In this paper we adopt two evaluation indexes commonly used in image retrieval: mean Average Precision (mAP) and Average Normalized Modified Retrieval Rate (ANMRR) [26].

Average Precision (AP) can be used as a global estimation measure to evaluate the retrieval performance. mAP (mean Average Precision) is the average of the AP of all queries, which has important reference value for evaluating the accuracy of retrieval results. The larger the mAP value is, the more accurate the result is [27].

The similarity ranking of images of retrieval results is also of great significance for evaluating retrieval performance. Average Normalized Modified Retrieval Rate (ANMRR) is based on Average Rank (AVR) and improved upon it. ANMRR can better reflect the retrieval performance by considering the order of returned images in the retrieval results. The value range of ANMRR is [0, 1]. When the retrieval result is the same type of image, ANMRR value is 0, and on the contrary, ANMRR value is 1. Therefore, when evaluating the model, the smaller the ANMRR is, the more correct images are ranked at the top, and the higher the accuracy is, the better the retrieval performance is.

3.2 Influence of Reduction Thresholds

To analyze the influence of the size of the retrieval region on the retrieval accuracy of image features, different reduction thresholds were set for the selected UC-Merced data-set. By fine-tuning ResNet50 model, we extracted the image features of retrieval regions of 0.5, 0.6, 0.7, 0.8, 0.9, 1.0 respectively, and then we conducted the image retrieval test directly. The results are shown in Table 3.

Table 3. Influence of different noise reduction thresholds

σ	0.5	0.6	0.7	0.8	0.9	1.0
mAP (%)	88.59	90.54	91.77	93.09	94.03	93.14
ANMRR	0.0870	0.0710	0.0617	0.0508	0.0448	0.0515

In the Table 3, we could found that when the threshold σ is set to be 0.5, mAP is the lowest, 88.59%, while ANMRR is the highest, 0.0870. And the retrieval effect of image features in the retrieval region is the worst. According to the principle of selecting value of threshold σ, the size of the retrieval area of this threshold would be reduced twice as much as that of the original image, and the features of the main object of the image would be partially lost. Therefore, all the information of the image can not be fully expressed, especially for the larger categories of ground object. When σ is set to be 1.0, the size of the retrieval region is the original image size, so the extracted features of the retrieval region are the original image features. When σ is set to be 0.9, the retrieval accuracy of mAP of original image increases from 93.14% to 94.03%, adding by 0.96%, while ANMRR decreases from 0.0515 to 0.0448, reducing by 13.01%. The retrieval accuracy obtained by this threshold 0.9 is the best among the other thresholds. Under this threshold 0.9, the main object of RS image can be effectively detected and divided into the designated retrieval region, so the extracted retrieval region features under the threshold which is set to be 0.9 can better express RS image than other thresholds. Therefore, the reduction threshold was set as 0.9 on the next experiments.

3.3 PCA Dimension Reduction Process

However, due to the vector dimension of the extracted basic retrieval features above is too high, we use PCA method to reduce the dimension of features, for reducing the retrieval time and avoiding the impact of redundant information between feature vectors. In order to analyze the influence of principal component dimension of extracted features on retrieval features, we take the matric es composed of unit feature vectors of the K rows before, where K is set to be 8, 16, 32, 64, 128, 256 respectively. Then Principal component analysis is performed on these matrices to obtain the corresponding image feature vectors and then the retrieval results were compared. Figure 4 shows the retrieval results of principal component features of unit feature vectors in the K rows before. From the Figure 4, we can see that when K is set to be 16 and 32, the mAP is 0.9497 and 0.9436, and ANMRR is 0.0358 and 0.0419, respectively. Compared with other K rows before, the feature retrieval effect of that is the best. Since K value has a linear relationship with the retrieval accuracy, it can be inferred that there is an optimal K value in the interval [16, 32], The interpolation method is used to select the optimal K value between 16 and 32, and K is set to 9. The mAP of its retrieval feature is 0.9517, and the ANMRR is 0.0345, which is the optimal retrieval result we can obtained.

3.4 Visual Representation

For verifying the retrieval performance of the method proposed here, we compared the method proposed here with the method that directly uses ResNet model to extract features. We selected two categories in this paper: beach and sparse housing. Figure 5 shows the comparison of retrieval results.

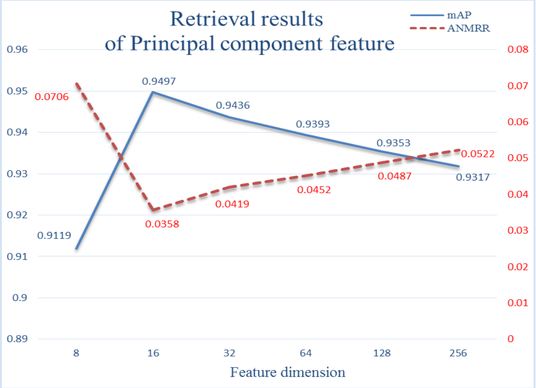

Figure 4. Principal component feature retrieval results of unit feature vectors in the K rows before

In Figure 5(a) and Figure 5(b), only 5 related images were extracted from the retrieval results of the previous behavior directly using features. In the next act, 7 and 8 related images are extracted respectively using the results obtained by the method proposed here. So the method in this paper can improve the retrieval effect.

3.5 Analysis of Category Results

For analyzing the effectiveness of class retrieval performance, we extract the retrieval features of each class of UCMD original image directly by fine-tuned the ResNet50 model, and obtained image retrieval results and then the retrieval results were compared with those of the proposed method. Table 4 shows the comparison of mAP and ANMRR of the retrieval features of each category of RS image respectively. Org_pic is the retrieval result of each category of original image, and KeyR_pic is the retrieval result of each category of regional image.

In Table 5, the mAP of the majority categories has raised by using the method in this paper. Especially, the retrieval accuracy of some categories, e.g., Medium residential, Highway, Parking lots, and Overpass have been raised to 100%. For the other categories such as Sparse residential and Beach, the mAP values of which were below 80%, we retrieved them difficult in the original image. While by using our method, the mAP of the Sparse residential is increased by about 8.96%, and that of Beach by 15.48%. The average value of all categories has also raised. Hence, the method proposed in this paper can obtain better results.

ANMRR values of all categories in this paper were also decreased to a certain degree. The ANMRR values for the Dense residential, Medium residential, Highway, Overpass, Mobile home park and Parking lot categories all fell to the zero. And the ANMRR of Beach, Intersection and River are all decreased by more than 55%, and the ANMRR values of Intersection category is the category with the biggest decline, reducing by about 61.33%. The ANMRR of the Chaparral and Golf course were increased slightly. In both of datasets, the ANMRR of the keyRegion is about 27.46% lower than original. Therefore, the method proposed can obtain better retrieval results of the same category.

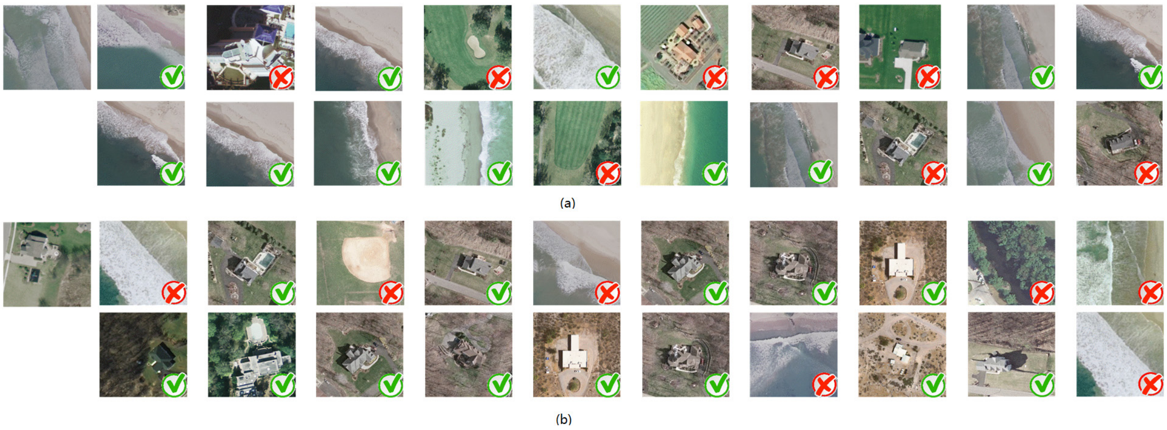

Figure 5. Comparison of two examples with different methods: (a) Beach (b) Sparse housing

3.6 Comparison with Other Methods

In this paper, the method proposed would compare with the other methods on UCMD datasets. The results of RS image retrieval methods of middle-level feature and high-level feature are compared respectively, shown as in Table 3. Among them, BoVW, VLAD and LSL are commonly used retrieval methods in the early stage, but it requires too much manpower, time, and the features extracted from middle-level feature cannot effectively express RS images. The ANMRR value is at a high level, but the category error rate of the retrieved image is severe. YOLOv5_ResNet50 in Table 4 is the result of image feature retrieval based on YOLOv5. RS image retrieval region were proposed and directly extracted by fine-tuning ResNet50 model. YOLOv5_ResNet50_PCA comes from the retrieval result of YOLOv5_ResNet50 reduced into 19 dimensional features excuated by PCAs reducing-dimension solutions. YOLOv5_ResNet50_PCA_W is the retrieval result of YOLOv5_ResNet50_PCA using class-based weighted distance method in the retrieval process. As shown in the Table 5, the retrieval results of YOLOv5_ResNet50_PCA, YOLOv5_ResNet50_PCA_W methods in this paper are better than those of previous methods with high-level features. In comparison, the final retrieval result mAP of this paper increased from 94.86% to 95.94%, an improvement of 1.14%. ANMRR decreased by 19.55% from 0.0404 to 0.0325. In the dimension of RS image retrieval feature, the method proposed here has reached the lowest 19 dimensions. Because the redundant information between feature vectors is removed, the retrieval operation is reduced, so the retrieval accuracy and efficiency are further improved. The above comparative analysis shows that the method proposed in this paper could effectively raise the performance of RS image retrieval. Moreover, it has been demonstrated that the application of PCAs algorithm always decrease ANMRR. The final mAPs are the most optimal data we collected. Generally, the YOLO structures in practical application are more easier using and less time processing. Once the detected target is a relatively larger target, the detection effect is not stable.

Table 5. Performances comparison with state-of-the-art methods of UCMD dataset

	Method	Dimension	mAP (%)	ANMRR
Middle level feature	Aptoula [28]	62	-	0.5750
	BoVW [10, 32]	15000	-	0.5910
	VLAD [12]	16384	-	0.4604
	LSL [29]	2048	-	0.5556
	ARGMM [30]	-	-	0.5748
High level feature	VGG16-conv5-IFK [16]	102400	51.02	0.4070
	VGG16-fc [16]	4096	52.47	0.3940
	LDCNN [17]	30	-	0.4390
	RBCP (FT VGGM) [13]	4608	64.47	0.3380
	RBCP (FT VGG16) [13]	2048	59.24	0.2889
	GoogLeNet (FT)-MultiPatch [27]	1024	-	0.3140
	GoogLeNet (FT)-MultiPatch-PCA [27]	32	-	0.2850
	VGG16_GoogLeNet_max_r_PCA [31]	32	-	0.2620
	FT_VGGM_GoogLeNet_max_r [31]	4096	66.21	0.2688
	Fc7_W (VGG16) [25]	-	90.86	0.0673
	Pool5_W (ResNet50) [25]	-	94.86	0.0404
	YOLOv5_ResNet50	2048	94.03	0.0448
	YOLOv5_ResNet50_PCA	19	95.17	0.0345
	YOLOv5_ResNet50_PCA_W	**19**	**95.94**	**0.0325**

4 Conclusions

In this paper, we present an RSIR algorithm based on CNN and critical region detection. For reducing the image background's interference, we firstly determined the key regions by YOLOv5 model. Then, ResNet was used to extracted the features of RS images, but due to the high dimension of features extracted, we used PCA method to reduce its dimension. Finally, the Class-based weighted distance method was used as the similarity measure criterion to further improve the retrieval performance. This method not only fully absorbs the object detection advantages of YOLO model, but also makes use of the feature extraction efficiency of ResNet. The results show that this method is superior to the existing RS image retrieval methods and effectively improves the retrieval performance.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 42174055, No. 41261091).

Table 4. Class mAP and ANMRR comparison of different methods

Class	mAP (%)		ANMRR		Class	mAP (%)		ANMRR	
	Org_pic	KeyR_pic	Org_pic	KeyR_pic		Org_pic	KeyR_pic	Org_pic	KeyR_pic
Argiculture	90.95	**93.98**	0.0694	**0.0563**	Intersection	87.19	**95.3**	0.1218	**0.0471**
Airplane	100	100	0	**0**	Medium residential	97.26	**100**	0.0256	**0**
Baseball diamond	94.8	**96.68**	0.0524	**0.0227**	Mobile home park	99.97	**100**	0.00014006	**0**
Beach	70.97	**86.45**	0.1862	**0.0831**	Overpass	99.42	**100**	0.0041	**0**
Buildings	95.78	**98.74**	0.0251	**0.0069**	Parking lot	99.98	**100**	0.00011204	**0**
Chaparral	98.32	96.58	0.0106	0.0255	River	80.85	**91.76**	0.1705	**0.0689**
Dense residential	99.99	**100**	0.000056022	**0**	Runway	100	100	0	**0**
Forest	94.37	**97.76**	0.0326	**0.0118**	Sparse residential	64	**72.96**	0.2607	**0.2328**
Highway	95.97	**100**	0.028	**0**	Storage tanks	98.68	**99.86**	0.0075	**0.00070028**
Golf course	89.97	88.31	0.0733	0.1025	Tennis court	97.68	**99.85**	0.012	**0.00084034**
Harbor	99.84	96.61	0.0012	0.0244	Average	93.14	**95.94**	0.0515	**0.0325**

References

[1] D. Li, S. Wang, H. Yuan, D. Li, Software and applications of spatial data mining, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Vol. 6, No. 3, pp. 84-114, May/ June, 2016.

[2] F. Ye, M. Dong, W. Luo, X. Chen, W. Min, A new re-ranking method based on convolutional neural network and two image-to-class distances for remote sensing image retrieval, *IEEE Access*, Vol. 7, pp. 141 498-141507, September, 2019.

[3] L. Fan, H. Zhao, H. Zhao, Distribution consistency loss for large scale remote sensing image retrieval, *Remote Sensing*, Vol. 12, No. 1, Article No. 175, January, 2020.

[4] L. Liu, Y. Wang, J. Peng, A. Plaza, DFLLR: Deep feature learning with latent relationship embedding for remote sensing image retrieval, *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 60, pp. 1-14, 2022.

[5] Y. Lv, W. Xiong, X. Zhang, Y. Cui, Fusion-based correlation learning model for cross-modal remote sensing image retrieval, *IEEE Geoscience and Remote Sensing Letters*, Vol. 19, pp. 1-5, 2022.

[6] Y. Ge, Z. Yang, Z. Huang, F. Ye, A multi-level feature fusion method based on pooling and similarity for hrrs image retrieval, *Remote Sensing Letters*, Vol. 12, No. 11, pp. 1090-1099, August, 2021.

[7] P. Bosilj, E. Aptoula, S. Lefèvre, E. Kijak, Retrieval of remote sensing images with pattern spectra descriptors, *ISPRS International Journal of Geo-Information*, Vol. 5, No. 12, Article No. 228, December, 2016.

[8] P. Napoletano, Visual descriptors for content-based retrieval of remote-sensing images, *International journal of remote sensing*, Vol. 39, No. 5, pp. 1343-1376, 2018.

[9] B. Zhang, X. Sun, L. Gao, L. Yang, Endmember extraction of hyperspectral remote sensing images based on the ant colony optimization (aco) algorithm, *IEEE transactions on geoscience and remote sensing*, Vol. 49, No. 7, pp. 2635-2646, July, 2011.

[10] Y. Yang, S. Newsam, Geographic image retrieval using local invariant features, *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 51, No. 2, pp. 818-832, February, 2013.

[11] J. Yang, J. Liu, Q. Dai, An improved bag-of-words framework for remote sensing image retrieval in large-scale image databases, *International Journal of Digital Earth*, Vol. 8, No. 4, pp. 273-292, 2015.

[12] S. Ozkan, T. Ateş, E. Tola, M. Soysal, E. Esen, Performance analysis of state-of-the-art representation methods for geographical image retrieval and categorization, *IEEE Geoscience and Remote Sensing Letters*, Vol. 11, No. 11, pp. 1996-2000, November, 2014.

[13] Y. Ge, Y. Tang, S. Jiang, L. Leng, S. Xu, F. Ye, Region-based cascade pooling of convolutional features for hrrs image retrieval, *Remote sensing letters*, Vol. 9, No. 10, pp. 1002-1010, August, 2018.

[14] F. Ye, Y. Su, H. Xiao, X. Zhao, W. Min, Remote sensing image registration using convolutional neural network

features, *IEEE Geoscience and Remote Sensing Letters*, Vol. 15, No. 2, pp. 232-236, February, 2018.

[15] Y. Ge, S. Jiang, Q. Xu, C. Jiang, F. Ye, Exploiting representations from pre-trained convolutional neural networks for high-resolution remote sensing image retrieval, *Multimedia Tools and Applications*, Vol. 77, No. 13, pp. 17489-17515, July, 2018.

[16] W. Zhou, S. Newsam, C. Li, Z. Shao, Patternnet: A benchmark dataset for performance evaluation of remote sensing image retrieval, *ISPRS journal of photogrammetry and remote sensing*, Vol. 145, pp. 197-209, November, 2018.

[17] W. Zhou, S. Newsam, C. Li, Z. Shao, Learning low dimensional convolutional neural networks for high-resolution remote sensing image retrieval, *Remote Sensing*, Vol. 9, No. 5, Article No. 489, May, 2017.

[18] W. Cui, F. Wang, X. He, D. Zhang, X. Xu, M. Yao, Z. Wang, J. Huang, Multi-scale semantic segmentation and spatial relationship recognition of remote sensing images based on an attention model, *Remote Sensing*, Vol. 11, No. 9, Article No. 1044, May, 2019.

[19] H. Li, H. Huang, L. Chen, J. Peng, H. Huang, Z. Cui, X. Mei, G. Wu, Adversarial Examples for CNN-Based SAR Image Classification: An Experience Study, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 14, pp. 1333-1347, February, 2021.

[20] Y. Li, Y. Zhang, X. Huang, J. Ma, Learning source-invariant deep hashing convolutional neural networks for cross-source remote sensing image retrieval, *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 56, No. 11, pp. 6521-6536, November, 2018.

[21] K. Kanwal, K. T. Ahmad, R. Khan, A. T. Abbasi, J. Li, Deep learning using symmetry, fast scores, shape-based filtering and spatial mapping integrated with cnn for large scale image retrieval, *Symmetry*, Vol. 12, No. 4, Article No. 612, April, 2020.

[22] P. Desai, J. Pujari, C. Sujatha, A. Kamble, A. Kambli, Hybrid approach for content-based image retrieval using vgg16 layered architecture and svm: An application of deep learning, *SN Computer Science*, Vol. 2, No. 3, pp. 1-9, May, 2021.

[23] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, Nevada, USA, 2016, pp. 770-778.

[24] Z. Zhuo, Z. Zhou, Remote Sensing Image Retrieval with Gabor-CA-ResNet and Split-Based Deep Feature Transform Network, *Remote Sensing*, Vol. 13, No. 5, Article No. 869, March, 2021.

[25] F. Ye, H. Xiao, X. Zhao, M. Dong, W. Luo, 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, October, 2018.

[26] B. S. Manjunath, J.-R. Ohm, V. V. Vasudevan, A. Yamada, Color and texture descriptors, *IEEE Transactions on circuits and systems for video technology*, Vol. 11, No. 6, pp. 703-715, June, 2001.

[27] X.-Y. Tong, G.-S. Xia, F. Hu, Y. Zhong, M. Datcu, L. Zhang, Exploiting deep features for remote sensing image retrieval: A systematic investigation, *IEEE Transactions on Big Data*, Vol. 6, No. 3, pp. 507-521, September, 2019.

[28] E. Aptoula, Remote sensing image retrieval with global morphological texture descriptors, *IEEE transactions on geoscience and remote sensing*, Vol. 52, No. 5, pp. 3023-3034, May, 2014.

[29] Z. Du, X. Li, X. Lu, Local structure learning in high resolution remote sensing image retrieval, *Neurocomputing*, Vol. 207, pp. 813-822, September, 2016.

[30] B. Chaudhuri, B. Demir, L. Bruzzone, S. Chaudhuri, 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, July, 2016.

[31] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, F. Fraundorfer, Deep learning in remote sensing: A comprehensive review and list of resources, *IEEE Geoscience and Remote Sensing Magazine*, Vol. 5, No. 4, pp. 8-36, December, 2017.

[32] Y. Yang, S. Newsam, Bag-of-Visual-Words and Spatial Extensions for Land-Use Classification, *The 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, San Jose, California, USA, 2010, pp. 270-279.

Biographies

Junwei Xin, Male, born in April 1987, is a doctoral candidate in geological resources and geological engineering of East China University of Technology, mainly engaged in the research on the integration of deep learning methods and geological remote sensing technology.

Famao Ye, Male, born in September 1978, doctor, associate professor of East China University of Technology, mainly engaged in remote sensing image processing, deep learning and artificial intelligence research.

Yuanping Xia, Male, born in April 1982, doctor, professor of East China University of Technology, mainly engaged in geological disaster investigation and early warning, remote sensing image processing and other research work.

Yan Luo, Female, born in October 1983, university lecturer, master, teacher of East China University of Technology, mainly engaged in the research of photogrammetry technology and remote sensing image processing methods.

Xiaoyong Chen, Male, born in September 1961, professor, doctor, doctoral supervisor, mainly engaged in theoretical research and application technology development of geographic information science.