Neurocomputing 207 (2016) 813–822

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Local structure learning in high resolution remote sensing image retrieval

Zhongxiang Du [a,b], Xuelong Li [a,*], Xiaoqiang Lu [a]

[a] Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, PR China
[b] University of the Chinese Academy of Sciences, 19A Yuquanlu, Beijing 100049, PR China

ARTICLE INFO

Article history:
Received 27 June 2015
Received in revised form
11 April 2016
Accepted 28 May 2016
Communicated by Zidong Wang
Available online 7 June 2016

Keywords:
High resolution remote sensing image retrieval
Lipschitz smooth
Manifold structure

ABSTRACT

High resolution remote sensing image captured by the satellites or the aircraft is of great help for military and civilian applications. In recent years, with an increasing amount of high resolution remote sensing images, it becomes more and more urgent to find a way to retrieve them. In this case, a few methods based on the statistical information of the local features are proposed, which have achieved good performances. However, most of the methods do not take the topological structure of the features into account. In this paper, we propose a new method to represent these images, by taking the structural information into consideration. The main contributions of this paper include: (1) mapping the features into a manifold space by a Lipschitz smooth function to enhance the representation ability of the features; (2) training an anchor set with several regularization constrains to get the intrinsic manifold structure. In the experiments, the method is applied to two challenging remote sensing image datasets: UC Merced land use dataset and Sydney dataset. Compared to the state-of-the-art approaches, the proposed method can achieve a more robust and commendable performance.

1. Introduction

Image retrieval is one of the most challenging tasks in the *high resolution remote sensing image* (HRRSI) analysis [1]. With the development of remote sensing imaging, an increasing number of high resolution images have been captured from the satellites or the aircrafts, which can provide abundant spatial and contextual information. In this case, retrieving these images efficiently and effectively has become one of the best interested research problems.

As a better representation can measure the image similarity more accurately [1], to address the retrieval problem, one of the most widely used strategies is to find an effective representation of the images. Image representation methods can be divided into two main streams. The first stream is supervised methods. This kind of methods needs a number of training samples to establish a discrimination model to obtain the image representation. However, with a large amount of different land cover types, a huge number of training samples are needed, which makes these methods unpractical in a real high resolution remote sensing image retrieval task. The second stream is the unsupervised methods. In most cases, the unsupervised methods are based on the local descriptors. With the development of the local descriptors, a series of robust features have been proposed, such as *scale-invariant feature transform* (SIFT) [2,3] and *speeded up robust features* (SURF) [4]. Based on these local descriptors, recently, some image representation techniques, including *bag of words* (BoW) [5,6], *fisher vector* (FV) [7] and *vector of locally aggregated descriptors* (VLAD) [8,9], have been proposed to obtain better results in remote sensing image retrieval. Nevertheless, these techniques have some drawbacks. In fact, retrieving remote sensing images, similar to the query image, relies on the capability and effectiveness of the feature representation. Researches have shown that high-dimensional Euclidean space cannot be uniformly filled up by the image data. Thus, the image data can be regarded as sampled data from or near a sub-manifold of an ambient space [10]. Therefore, it is necessary to consider the intrinsic structure in the high-dimensional space, while representing and describing high resolution images.

In this paper, we propose an unsupervised image retrieval method, named as *local structure learning* (LSL), to represent and describe high resolution images. Unlike the aforementioned methods, the proposed method aims at describing the data by learning the corresponding representation in the high-dimensional space. In order to exploit the intrinsic structural information of the original data, LSL incorporates graph regularization to learn

* Corresponding author.
E-mail address: xuelong_li@opt.ac.cn (X. Li).

http://dx.doi.org/10.1016/j.neucom.2016.05.061
0925-2312/

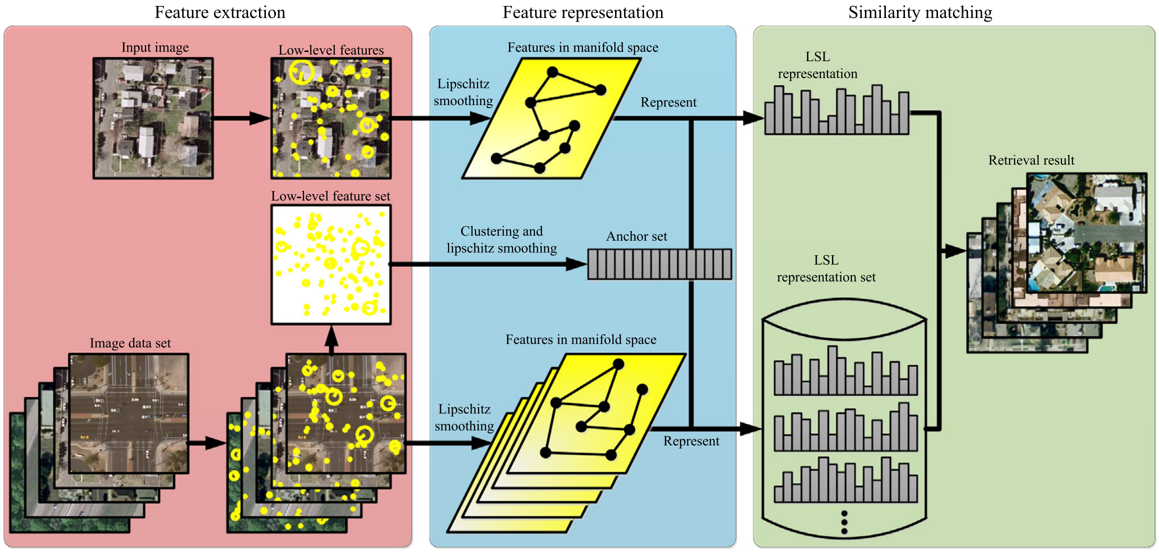

Fig. 1. The framework of the method proposed in this paper.

the representation. In this case, the LSL can capture or reflect the characteristics of the original data. Similar to BoW, VLAD and FV, by calculating the inner product of two LSL representations, we can get the similarity of two corresponding images.

The proposed method can be divided into three steps. First, a Lipschitz smooth function [11] is introduced to map the original space to a high-dimensional space. In this case, the data in the high dimensional space has a more powerful representation ability. Second, an anchor set is constructed in the high dimensional space for a better representation, and a nearest neighbor graph is built to model the local structural information of the original data. Finally, represent the mapped features with the anchor set. Therefore, it can achieve a more robust performance than the traditional representations.

The framework of the proposed method in this paper is shown in Fig. 1, which describes the complete workflow of a real HRRSI retrieval system.

The rest of this paper is organized as follows. Section 2 gives a brief introduction about the general methods on high resolution remote sensing image retrieval. Section 3 provides a detailed description of the method proposed in this work. In Section 4, the measure of performance and an analysis of the experiment results on two *land-use/land-cover* (LULC) datasets are given. Section 5 concludes this paper.

2. Related works

In this section, several works on the high resolution remote sensing image retrieval are introduced, including methods based on the statistical information of raw pixels, methods based on the statistical information of low-level features, and the relevance feedback methods.

Methods based on the statistical information of raw pixels: The statistical information of raw pixels can be divided into color histogram features [12] and texture features [13]. Color histogram features can be directly computed in the *red green blue* (RGB) color space. Sometimes *hue saturation value* (HSV) or *hue saturation lightness* (HSL) [14] color space can also be adopted, as they are more intuitive and perceptually relevant. Texture feature is a little more complex. Many methods can be used to get this feature, and the *local binary patterns* (LBP) [15] is one of the most widely utilized methods. Some other methods based on the Gabor filters [16] are also proposed to get the texture feature. By using the Gabor filters to turn the original image into different scales and orientations to form a set of new images, expectation and variance of the new images can be calculated, then a vector is formed to describe the original image.

Methods based on the statistical information of low-level features: With the development of the feature representation, some features, such as SIFT and SURF, are shown to have the ability to handle intensity, rotation, scale and affine variations to some extent. These low-level features can express the stable patterns in images. So, compared to the statistical information of raw pixels, the information of these features is more robust [17–19]. Then, to obtain the statistical information of these features, the BoW [20], FV and VLAD [21] are proposed. The BoW model is firstly introduced on the document classification and retrieval. In image retrieval, these stable features can be treated as the words in a document. There is a codebook in the BoW model. To generate this codebook, one of the simplest methods is performing *k-means* [22] clustering over all the feature vectors. Finally every image in the dataset can be represented as a k-dimensional vector.

The FV model is based on the fisher kernel [23]. The pattern classification can be divided into generative model and discrimination model, and the fisher kernel connects the two models together. The *probability density function* (PDF) of the features is denoted as p, and its parameter is λ. After normalized by the fisher information matrix F_λ, a feature set $X = \{x_i, i = 1\ldots T\}$ can be represented as the following equation:

$$G_\lambda^X = \sum_{t=1}^{T} F_\lambda^{-1/2} \nabla_\lambda \log p(x_t|\lambda). \tag{1}$$

To obtain the parameter λ, a *Gaussian mixture model* (GMM) [24,25] is adopted. Assuming features of the images subject to the Gaussian distribution, K Gaussian distributions are applied to model the features. So the parameter λ can be denoted as $\lambda = \{w_k, \mu_k, \sigma_k.\ k = 1\ldots K\}$. When setting λ into w_k, μ_k and σ_k separately, we can obtain the zero-order, first-order and second-order statistical features. By jointing the three different features together, the final FV with the PDF can be obtained by GMM. Suppose D-dimensional local feature is used, thus the length of this vector is $(2 \times D + 1) \times K - 1$. In order to reduce its dimensionality, *product quantization* (PQ) [26] or *principal component analysis* (PCA) is usually applied.

VLAD can be regarded as a simplification of FV, while its