ISPRS Journal of Photogrammetry and Remote Sensing 144 (2018) 202–216

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

Projection learning with local and global consistency constraints for scene classification

Panpan Zhu[a], Liqiang Zhang[a,*], Yuebin Wang[a], Jie Mei[a], Guoqing Zhou[b,*], Fangyu Liu[a], Weiwei Liu[c], P. Takis Mathiopoulos[d]

[a] Beijing Key Laboratory of Environmental Remote Sensing and Digital Cities, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China
[b] College of Geomatics and Geoinformation, Guilin University of Technology, Guilin 541004, China
[c] Beijing Obstetrics and Gynecology Hospital, Capital Medical University, Beijing 100006, China
[d] Department of Informatics and Telecommunications, National and Kapodestrian University of Athens, Athens 15784, Greece

ARTICLE INFO

Keywords:
Scene classification
Projection learning
Manifold regularization
Label consistency
Classification error

ABSTRACT

Besides the problem of high within-class variation and between-class ambiguity in high spatial resolution (HSR) remote sensing images, the dimension of data representation is very high, which poses a challenge for scene classification. To achieve high scene classification performance, it is important to uncover a discriminative subspace for data representation and scene classification. In this paper, we propose a projection learning framework with local and global consistency constraints for aerial scene classification. During the learning process, the within-class compactness and between-class separation of the data representation are enforced. To guarantee the subspace to be locally smooth, we obtain the local geometric structures of data including the similarity of features and geospatial adjacency of image patches. We utilize the global label consistency constraint to enforce the discrimination of the subspace. To make the projection optimal for classification, the projection learning with local and global constraints is integrated with the classification error to form a unified objective function. An efficient iteration algorithm is employed to solve the objective function. Experimental results demonstrate the superior performance of our method over state-of-the-art algorithms on aerial scene classification tasks.

1. Introduction

With increasing number of high spatial resolution (HSR) remote-sensing images, scene classification has been an active research topic (Lu and Weng, 2007; Myint et al., 2011; Ge et al., 2016), which aims to label an image with a specific semantic category (Zhang et al., 2011a, 2011b; Xu et al., 2016). It remains challenging since remote sensing images are with high within-class variation and between-class ambiguity (Li et al., 2011; Liu et al., 2014; Ma et al., 2017; Mui et al., 2015; Hu et al., 2015; Zhu et al., 2017; Marmanis et al., 2018).

Scene-level HSR image classification plays an important role in remote sensing images understanding. Many works have been reported about scene classification tasks (Gerke and Xiao, 2014; Huang et al., 2018; Zhang et al., 2018; Ge et al., 2014). Among these works, the bag-of-visual-words (BoVW) based methods (Sivic and Zisserman, 2003; Wang et al., 2010; Zhao et al., 2016) have been widely applied. To use the spatial information, the spatial pyramid matching (SPM) (Lazebnik et al., 2006) partitions the image into a set of sub-regions, and then compute the BoVW histogram of each subregion at all levels. It finally concatenates all the histograms to form a spatial pyramid representation of the image. To reduce the computational costs, Yang et al. (2009) proposed a linear SPM method using sparse coding to encode local descriptors.

Different from the approach of Yang et al. (2009), sparse coding methods use sparse coding to encode the extracted image features instead of the local descriptors. Many sparse coding methods are examined for image classification purposes (Aharon, 2006; Wright et al., 2009; Zhang and Li, 2010; Feng et al., 2017). In the sparse representation-based classification (SRC) (Wright et al., 2009), the sparsity is appropriately harnessed in addressing the problems of face recognition. Directly taking all training samples as the dictionary may result in a large and redundant dictionary. While the K-SVD algorithm (Aharon, 2006) learns the sparse representation coefficient and an overcomplete dictionary simultaneously. To optimize the feature representation and classifier simultaneously, Zhang and Li (2010) developed the discriminative K-SVD (D-KSVD) algorithm by explicitly incorporating the classification error into the optimal function. To use the supervised information, Jiang et al. (2013) integrated a label

* Corresponding authors.
E-mail address: zhanglq@bnu.edu.cn (L. Zhang).

https://doi.org/10.1016/j.isprsjprs.2018.07.004
Received 14 December 2017; Received in revised form 29 May 2018; Accepted 7 July 2018
0924-2716/

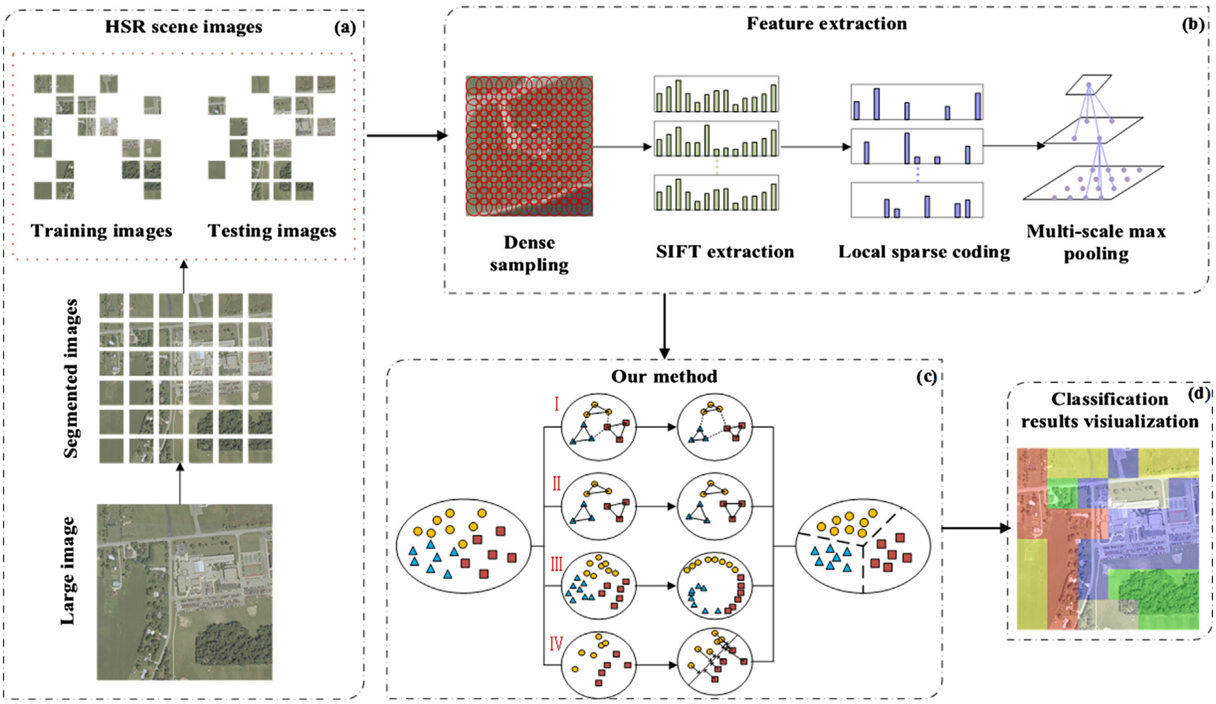

Fig. 1. The framework of the proposed method. (a) HSR scene images. (b) Feature extraction procedure. (c) Projection learning process. I, II, III, and IV corresponds to the projection learning, label consistency, co-graph regularization and classification error, respectively. (d) Classification results.

consistency term into the objective function and obtained high-quality classification results. In Feng and Zhou (2016), the correlation structure is introduced into the SRC classifier to make the estimation of the sparse coefficient matrix more stable. Based on the SRC classifier, Feng and Zhou (2017) further proposed a kernel regularized data uncertainty classifier that utilized the Tikhonov matrix to encode the importance of each sample.

Recently, deep learning has attracted considerable interests (Dean et al., 2012; Zhao et al., 2017; Zhang and Zhang, 2018). Krizhevsky et al. (2012) developed a deep convolutional neural network to classify high-resolution images. A discriminant deep belief network (Zhao et al., 2017) was proposed to learn high-level features for synthetic aperture radar image classification. However, deep learning methods usually need large training samples and consume high computational costs.

A common disadvantage of the above-mentioned methods (Sivic and Zisserman, 2003; Lazebnik et al., 2006; Yang et al., 2009; Jiang et al., 2013) is that the dimension of data representation is large. High-dimensional data representation often introduces noise and redundant information (Li et al., 2015). What is worse, the computational efficiency drops exponentially as the dimensionality increases. Therefore, it is a fundamental problem to uncover a low-dimensional representation of high- dimensional data.

To uncover a low-dimensional representation of high-dimensional data, many approaches rely on projection learning (Wright et al., 2009; Wang et al., 2015a, 2015b) as it maps original high-dimensional data representation into low-dimensional subspaces with the desired statistical or geometric properties well being preserved. A variety of projection learning algorithms, including principal component analysis (PCA) (Martínez and Kak, 2001), linear discriminant analysis (LDA) (Yu and Yang, 2001), locality preserving projections (He and Niyogi, 2003), and optimal mean robust PCA (Nie et al., 2014), use a projection matrix $\mathbf{A}$ to map the data matrix $\mathbf{X}$ in the original feature space to a low-dimensional space $\mathbf{A}^T\mathbf{X}$. The learned projections of these methods are a linear combination of original features. For example, Wang and Zhang (2007) proposed a linear feature extraction method in a supervised manner to learn a projection matrix by narrowing the within-class distance and enlarging the between-class separation. Feng et al. (2016) tried to minimize the related pairwise subspaces and at the same time maximized the unrelated pairwise subspaces. Some manifold learning algorithms are studied to uncover the underlying nonlinear subspace, such as ISOMAP (Tenenbaum et al., 2000), linear embedding (Belkin and Niyogi, 2001) and locally linear embedding (Roweis and Saul, 2000). These methods look for hidden subspaces that optimally preserve local intrinsic manifold structure of data representation. Yan et al. (2007) presented a common graph embedding framework to unify a family of projection learning algorithms and indicated that the differences between these algorithms lie in the selection of the constraint and the use of similarity criteria. The similarity criteria can be Gaussian similarity from Euclidean distance (Belkin and Niyogi, 2001), local neighborhood relationship (Wang et al., 2015a, 2015b; Roweis and Saul, 2000), and prior class label information in supervised learning algorithms (Yu and Yang, 2001). In the general tensor discriminant analysis (Tao et al., 2007) and supervised tensor learning (Tao et al., 2005), the discriminative information is preserved during the training stage by leveraging the class label information. To guarantee the learned subspace to be compact and discriminative, some joint learning frameworks were proposed. Li et al. (2015) proposed a robust structured subspace learning method to uncover an appropriate latent subspace for data representation. The idea behind their method is to integrate the intrinsic geometric structure of data and the local and global structural consistencies of labels into a framework in order to guarantee the compactness and discrimination of the subspace. To obtain the nonnegative low-dimensional representation of data for image classification, Lu et al. (2017) proposed a subspace learning method called nonnegative discriminant matrix factorization (NDMF). NDMF projects the low-dimensional representation of the subspace of the base matrix to regularize the discriminant matrix factorization for discriminant subspace learning. Lu et al. (2018) later combined the structurally incoherent learning and low-rank learning with neighborhood preserving projection to form a unified model for enhancing image feature representation.

To uncover a compact and discriminative subspace, we propose a joint projection learning framework with the local and global consistencies for scene classification. The overview of the framework is