Hessian-Regularized Multitask Dictionary Learning for Remote Sensing Image Recognition

Guanhua Feng, Weifeng Liu, Shuying Li, Dapeng Tao, and Yicong Zhou

***Abstract*— Learning effective image representations is a vital issue for remote sensing (RS) image recognition tasks. Although numerous algorithms have been proposed, it is still challenging due to the limited labeled data. One representative work is the Laplacian-regularized multitask dictionary learning (LR-MTDL) that employs graph Laplacian regularization terms to fully utilize both the labeled and unlabeled information. However, it probably conduces to poor extrapolating power because Laplacian regularization biases the solution toward a constant function. In this letter, we propose a Hessian-regularized multitask dictionary learning to learn a source-data set-shared but target-data set-biased representation for RS image recognition. Particularly, Hessian can properly exploit the intrinsic local geometry of the data manifold and finally leverage the performance. Extensive experiments on four RS image data sets validate the effectiveness of the proposed method by comparing with baseline algorithms including single-task dictionary learning and LR-MTDL.**

***Index Terms*— Hessian regularization, multitask dictionary learning, remote sensing (RS) image recognition.**

I. Introduction

INSPIRED by the rapid progress of satellite and remote sensing (RS) technology, a huge quantity of RS images has been available nowadays. These images always contain sufficient information of space and spectra, which are of vital significance for earth observation applications, such as object detecting, traffic management, and urban planning. However, due to various geometrical structures and intricate spatial patterns, learning effective representations from RS data and realizing recognition is still a challenging task which has attracted great attention in the RS field. For the sake of recognizing and analyzing scenes from RS images, a volume of scene classification algorithms has been introduced and these algorithms can be divided into the following five categories: methods based on hand-craft features, methods based on unsupervised feature learning, methods based on deep learning, methods based on transfer learning, and methods based on manifold learning.

A. Methods Based on Hand-Craft Features

Methods based on hand-craft features utilize various hand-craft local image descriptors to represent images. Dos Santos *et al.* [1] developed a considerable study to explore several color and texture image descriptors including color histogram [2] and local binary pattern [3] for RS retrieval and classification. Sivic and Zisserman [4] proposed the bag-of-visual-words (BOVW) model to represent images with the frequency of visual words which are constructed by quantizing local features with a clustering approach. Afterward, several extensions of BOVW including the spatial pyramid model (SPM) [5] and cooccurrences-based SPM (SPM++) [6] were also introduced for recognizing the scenes.

B. Methods Based on Unsupervised Feature Learning

Methods based on unsupervised feature learning aim to automatically learn adaptive feature representations from images. Cheriyadat [7] employed a sparse coding-based method by encoding dense low-level feature descriptors in terms of basic functions to establish holistic sparse representation for aerial images. Luo *et al.* [8] proposed a large margin multimodel multitask feature extraction framework that is effective for learning strongly predictive feature representation. Zhang *et al.* [9] used the sparse autoencoder framework to extract the features of image patches by exploiting the local structural and spatial information.

C. Methods Based on Deep Learning

Methods based on deep learning try to adaptively learn effective image features with a multistage global feature learning architecture. Penatti *et al.* [10] directly extracted deep features from pretrained CaffeNet into aerial scene classification. Castelluccio *et al.* [11] used GoogLeNet with fine-tuning on the target RS data set and achieved encouraging classification performance. Wang *et al.* [12] presented a novel recurrent attention structure for hyperspectral image classification by constructing an effective end-to-end network based on it.

Manuscript received August 11, 2018; revised November 5, 2018; accepted November 14, 2018. Date of publication December 6, 2018; date of current version April 22, 2019. This work was supported in part by the National Natural Science Foundation of China under Grant 61671480, in part by the Foundation of Shandong Province under Grant ZR2018MF017, in part by the Fundamental Research Funds for the Central Universities, China University of Petroleum (East China) under Grant 18CX07011A and Grant YCX2017059, in part by the National Natural Science Foundation of China under Grant 61772455 and Grant U1713213, in part by the Yunnan Natural Science Funds under Grant 2016FB105, in part by the Program for Excellent Young Talents of Yunnan University under Grant WX069051, in part by the Macau Science and Technology Development Fund under Grant FDCT/189/2017/A3, in part by the Research Committee at University of Macau under Grant MYRG2016-00123-FST and Grant MYRG2018-00136-FST, and in part by the National Natural Science Foundation of China under Grant 61701387. *(Corresponding authors: Weifeng Liu; Yicong Zhou.)*

G. Feng and W. Liu are with the School of Information and Control Engineering, China University of Petroleum (East China), Qingdao 266580, China (e-mail: axjlgh@163.com; liuwf@upc.edu.cn).

S. Li is with the 16th Institute, China Aerospace Science and Technology Corporation, Xi'an 710100, China (e-mail: angle_lisy@163.com).

D. Tao is with the School of Information Science and Engineering, Yunnan University, Kunming 650091, China (e-mail: dapeng.tao@gmail.com).

Y. Zhou is with the Faculty of Science and Technology, University of Macau, Macau 999078, China (e-mail: yicongzhou@umac.mo).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LGRS.2018.2881834

D. Methods Based on Transfer Learning

Methods based on transfer learning take the advantage of reusing knowledge or information from source data to solve the classification problem in the related target data. Matasci *et al.* [13] proposed a semisupervised transfer component analysis for hyperspectral images, which statistically aligns the target image to the source image through a nonlinear transformation. Luo *et al.* [14] developed a heterogeneous multitask metric learning framework to exploit high-order information and obtain reliable feature transformations and metrics.

E. Methods Based on Manifold Learning

Methods based on manifold learning always seek a low-dimensional subspace in which certain local geometric properties of the original features can be preserved. Wang *et al.* [15] presented locality constraint criterion and structure preserving method for hyperspectral image classification based on low-rank representation. Luo *et al.* [16] introduced multiview vector-valued manifold regularization to make use of the geometric and structural information of features. Wang *et al.* [17] took the manifold structure into consideration instead of rating the similarities in the target space to properly assess the hyperspectral data structure. Peng *et al.* [18] incorporated a graph Laplacian regularization into unsupervised multitask dictionary learning (LR-MTDL) framework which significantly leverages the performance.

Although LR-MTDL has achieved cracking performance, it has been proven that graph Laplacian biases the solution toward a constant function and lacks extrapolating power [19]. In order to address these problems, we present a Hessian-regularized multitask dictionary learning (HR-MTDL) in this letter. In contrast to graph Laplacian, Hessian does not only have a richer null space but also drive the solution varying smoothly along the underlying manifold. Therefore, Hessian regularization is better to encode the local geometry than Laplacian regularization. We carefully implement HR-MTDL for RS image classification and conduct experiments on the AID data set, UC-Merced data set, WHU-RS19 data set, and RSSCN7 data set. The experimental results verify the effectiveness of the proposed method by comparing with several baseline algorithms.

The rest of this letter is arranged as follows. Section II briefly reviews some related work. Section III provides the proposed HR-MTDL approach. Section IV describes the details of the algorithm. Section V reports the experimental results. Section VI concludes this letter.

II. Related Work

In this section, we first give a brief review of several related works of the proposed algorithm including dictionary learning and LR-MTDL framework. Afterward, we introduce Hessian regularization.

A. Dictionary Learning

Dictionary learning is a class of unsupervised methods for learning sets of overcomplete bases to represent data efficiently. Suppose we are given a set of training samples $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{d\times n}$. Dictionary learning aims to learn a dictionary $D \in \mathbb{R}^{d\times k}$ and the sparse code matrix $A = [\alpha_1, \alpha_2, \ldots, \alpha_n] \in \mathbb{R}^{k\times n}$ by solving the following optimization problem:

$$\min_{D\in\mathcal{D},\alpha_i\in\mathbb{R}^k} \sum_{i=1}^{n} \|X - DA\|_F^2 + \lambda\Omega(A) \tag{1}$$

where $\mathcal{D} \triangleq \left\{D \in \mathbb{R}^{d\times k} \ |\forall i, \|d_i\|_2 \le 1\right\}$, $\|\cdot\|_F$ is the Frobenius norm of matrix. $\Omega(A)$ is the regularization term for promoting the sparsity constraint and λ is a positive parameter. when ℓ_0-norm or ℓ_1-norm is employed as the regularization term, the optimization problem can be effectively solved with the K-singular value decomposition algorithm and online dictionary learning methods, respectively.

B. Laplacian Regularized Multitask Dictionary Learning

Multitask dictionary learning incorporates multitask learning which belongs to an inductive transfer mechanism into dictionary learning. In RS, one data set always includes multiple sets of samples and each set is collected at a specific geographic situation. Instead of building individual classifiers for each sensing task, it is better to share data across tasks, which means what is learned from one task is transferred to the other correlated task. By learning the tasks in parallel under a shared representation, the transfer of knowledge among tasks is exploited to benefit all. This process works particularly well in the situation when there are limited training data related with each task. By taking advantage of data from associated tasks, the training data from each task are strengthened and the generalization ability of classifier enhanced. Assume we are given a few tasks $X = [X_1, X_2, \ldots, X_k], k = 1, \cdots K$. One of these tasks is the target task and the others are source tasks. In the multitask dictionary learning model, it always decomposes the dictionary that to be learned into two parts: a shared dictionary that captures latent attributes between all the tasks and a task-specific dictionary that captures unique aspects of task. Based on these considerations, LR-MTDL framework was developed. The formulation of this model is expressed as follows:

$$\begin{aligned}
&\left[D^c, D_K^u, D_1^r, \ldots, D_K^r\right] \\
&= \operatorname{argmin} \sum_{k=1}^{K} \mu\left(\left\|X_k - D^c A_k^c\right\|_F^2 + \left\|X_k - D^c A_k^c - D_k^r A_k^r\right\|_F^2\right) \\
&\quad + \left\|X_K - D^c A_K^c\right\|_F^2 \\
&\quad + \left\|X_K - D^c A_K^c - D_K^u A_K^u\right\|_F^2 \\
&\quad + \left\|X_K - D^c A_K^c - D_K^u A_K^u - D_K^r A_K^r\right\|_F^2 \\
&\quad + \eta \sum_{k=1}^{K} \sum_{i,j=1}^{N_k} w_{k,i,j} \left\|\alpha_{k,i}^c - \alpha_{k,j}^c\right\|^2 \\
&\quad + \eta \sum_{i,j=1}^{N_k} w_{K,i,j} \left\|\alpha_{K,i}^u - \alpha_{K,j}^u\right\|^2 \\
&\quad \text{s.t. } \left\|d_i^c\right\|_2^2 \le 1, \quad \left\|d_{K,i}^u\right\|_2^2 \le 1, \ \left\|d_{k,i}^r\right\|_2^2 \le 1, \ \forall i, k
\end{aligned} \tag{2}$$

where D^c denotes the task-shared dictionary, and D_K^u and D_k^r denote the task-specific dictionaries. A_k^c, A_K^u, and A_k^r are the

codes corresponding to them, d_i^c, $d_{K,i}^u$, and $d_{k,i}^r$ denote the ith column of D^c, D_K^u, and D_k^r respectively, $\alpha_{k,i}^c$ and $\alpha_{K,i}^u$ represent the ith column of A_k^c and A_K^u, respectively, and μ and η are the parameters of cost function terms. The last two terms are regularization terms that can be rewritten in the form of Laplacian matrix

$$\sum_{i,j=1}^{N_k} w_{k,i,j} \left\| \alpha_{k,i}^c - \alpha_{k,j}^c \right\|^2 = \mathrm{tr}\left(A_k^c L_k A_k^{c\prime} \right) \tag{3}$$

where $L_k = D_k - W_k$ is known as the Laplacian matrix, and D_k is a diagonal matrix in which diagonal elements are equal to the sum of the row entries of W_k.

C. Hessian Regularization

Suppose $M \subset \mathbb{R}^m$ is a smooth data manifold, and the tangent space $T_{X_i}(M) \subset \mathbb{R}^m$ at each point $X_i \in M$ can be defined. By considering the tangent space as a subspace of $\mathbb{R}^n$, we think of such tangent space $T_{X_i}(M) \subset \mathbb{R}^m$ an orthonormal coordinate system by utilizing the inner product inherited from $\mathbb{R}^n$. Then, the Hessian of function f is defined by evaluating Eells energy $S_{\mathrm{Eells}}(f)$ which is written for real-valued function, $f : M \to \mathbb{R}$, as

$$S_{Eells}(f) = \int \|\nabla_a \nabla_b\|^2_{T_x^* M \otimes T_x^* M} dV(x) \tag{4}$$

where $\nabla_a \nabla_b$ is the second covariant derivative of f, $dV(x)$ is the volume element. Orthonormal coordinates at a given point X_i are coordinates on M such that the manifold looks as Euclidean as possible (up to second order) around X_i. Therefore, in orthonormal coordinates x_r centered at point X_i

$$\nabla_a \nabla_b |_{X_i} = \sum_{r,s=1}^{m} \left. \frac{\partial^2 f}{\partial_{x_r} \partial_{x_s}} \right|_{X_i} dx_a^r \otimes dx_b^s \tag{5}$$

$$\|\nabla_a \nabla_b\|^2_{T_x^* M \otimes T_x^* M} = \sum_{r,s=1}^{m} \left(\frac{\partial^2 f}{\partial_{x_r} \partial_{x_s}} \right)^2 = \sum_{\alpha,\beta=1}^{k} \mathbf{f}_\alpha \mathbf{f}_\beta B_{\alpha\beta}^{(i)} \tag{6}$$

where $B_{\alpha\beta}^{(i)} = \sum_{r,s=1}^{n} H_{rs\alpha}^{(i)} H_{rs\beta}^{(i)}$. H denotes the operator, $\mathbf{f} \in \mathbb{R}^k$, $\mathbf{f}_j = f(X_j)$, $X_j \in N_k(X_i)$. $N_k(X_i)$ denotes the set of k nearest neighbors of point X_i. Obviously, the norm of the second covariant derivative is just the Frobenius norm of the Hessian of f in orthonormal coordinates. Summing over all data points, the final norm is formulated as follows:

$$\begin{aligned} S_{rmHess}(f) &= \sum_{i=1}^{n} \sum_{r,s=1}^{m} \left(\frac{\partial^2 f}{\partial_{x_r} \partial_{x_s}} |_{X_i} \right)^2 \\ &= \sum_{i=1}^{n} \sum_{\alpha \in N_k(X_i)} \sum_{\beta \in N_k(X_i)} \mathbf{f}_\alpha \mathbf{f}_\beta B_{\alpha\beta}^{(i)} \\ &= \langle f, Bf \rangle. \end{aligned} \tag{7}$$

Here, we call B Hessian matrix and the resulting function $S_{\mathrm{Hess}}(f)$ Hessian regularization. It has been demonstrated in [20] that Hessian regularization has a richer null space than Laplacian regularization which the null space is the constant functions on M.

III. Hessian-Regularized Multitask Dictionary Learning

In this section, we introduce our proposed HR-MTDL. Suppose we are given some tasks $X = [X_1, X_2, \ldots, X_t], t = 1, \cdots T$, $X_t \in \mathbb{R}^{M \times k}$ is the feature matrix with each column $x_{t,i}$ associating to an M-dimensional vector in the task t which consists of N_t samples. Task T is the target task and the rest are source tasks. We simply divide our framework into two halves. In the first halve, we aim for learning a shared dictionary D^c by using all the tasks and two task-specific dictionaries D_T^u and D_t^r. The first dictionary D^c is used to encode the latent attributes shared by all the tasks. The second dictionary D_T^u is unique to the target task T. The third dictionary D_t^r is a task-specific residual dictionary for encoding the residual parts of features that cannot be captured by D^c or D_T^u. Thus, the first halve can be formulated as follows:

$$\begin{aligned} &L(X, D, A) \\ &= \sum_{t=1}^{T} \gamma \left(\left\| X_t - D^c A_t^c \right\|_F^2 + \left\| X_t - D^c A_t^c - D_t^r A_t^r \right\|_F^2 \right) \\ &\quad + \left\| X_T - D^c A_T^c \right\|_F^2 \\ &\quad + \left\| X_T - D^c A_T^c - D_T^u A_T^u \right\|_F^2 \\ &\quad + \left\| X_T - D^c A_T^c - D_T^u A_T^u - D_T^r A_T^r \right\|_F^2. \end{aligned} \tag{8}$$

The minimization of the first two reconstruction error terms guarantees that D^c and D_t^r can better encode X_t and the residual part of X_t, respectively. The last three terms apply the same reconstruction formulation into target task T. The second halve is the Hessian regularization which is formulated as follows:

$$\sum_{i,j=1}^{N_t} H_{t,i,j} \left\| \alpha_{t,i}^c - \alpha_{t,j}^c \right\|^2 = \mathrm{tr}\left(A_t^c B_t A_t^{c\prime} \right) \tag{9}$$

where $H_{t,i,j}$ is the weighted matrix that generates penalty and B is the Hessian matrix. Obviously, the norm of the second covariant derivative is just the Frobenius norm of the Hessian of f in orthonormal coordinates. Summing up the energy of all the points make sure that the squared norm of Hessian is weighted with local density of the points, which leads to a stronger penalization of Hessian in densely sampled regions. Combining the above-mentioned two terms, HR-MTDL has the following expression:

$$\begin{aligned} &\left[D^c, D_T^u, D_1^r, \ldots, D_T^r \right] \\ &= \mathrm{argmin} \sum_{t=1}^{T} \gamma \left(\left\| X_t - D^c A_t^c \right\|_F^2 + \left\| X_t - D^c A_t^c - D_t^r A_t^r \right\|_F^2 \right) \\ &\quad + \left\| X_T - D^c A_T^c \right\|_F^2 \\ &\quad + \left\| X_T - D^c A_T^c - D_T^u A_T^u \right\|_F^2 \\ &\quad + \left\| X_T - D^c A_T^c - D_T^u A_T^u - D_T^r A_T^r \right\|_F^2 \\ &\quad + \lambda \sum_{i,j=1}^{N_t} \mathrm{tr}\left(A_t^c B_t A_t^{c\prime} \right) + \lambda \mathrm{tr}\left(A_T^u B_T A_T^{u\prime} \right) \\ &\quad \text{s.t. } \left\| d_i^c \right\|_2^2 \le 1, \quad \left\| d_{T,i}^u \right\|_2^2 \le 1, \ \left\| d_{t,i}^r \right\|_2^2 \le 1, \ \forall i, t \end{aligned} \tag{10}$$

where A_t^c, A_T^u, and A_t^r are the learned sparse codes. d_i^c, $d_{T,i}^u$, and $d_{t,i}^r$ denote the ith column of D^c, D_T^u, and D_t^r, respectively, and γ and λ are the positive parameters.

IV. Algorithm

In this section, we provide the details of optimization procedure. First, we fix D^c, D_T^u, D_t^r, and A_t^r to compute A_t^c and A_T^u. Equation (10) can be rewritten as

$$\min \|\bar{X}_t - \bar{D}\bar{A}_t\|_F^2 + \lambda \mathrm{tr}\left(\bar{A}_t \bar{B}_T \bar{A}_t'\right) \tag{11}$$

when $t \neq T$

$$\begin{aligned} \bar{X}_t &= \left[\mu X_t, \mu\left(X_t - D_t^r A_t^r\right)\right]^{\mathrm{T}} \\ \bar{D} &= [\mu D^c, D^c]^{\mathrm{T}} \\ \bar{A}_t &= \left[A_t^c\right]^{\mathrm{T}} \end{aligned} \tag{12}$$

when $t = T$

$$\begin{aligned} \bar{X}_t &= \left[X_T, X_T, X_T - D_T^r A_T^r\right]^{\mathrm{T}} \\ \bar{D} &= \left[D^c\ 0, D^c\ D_T^u, D^c\ D_T^u\right]^{\mathrm{T}} \\ \bar{A}_t &= \left[A_t^c, A_T^u\right]^{\mathrm{T}}. \end{aligned} \tag{13}$$

Then, we can obtain $\bar{\alpha_{t,i}}$ which is the ith column of $\bar{A_t}$ by setting the derivative of (11) equals zero. Second, we fix other terms to compute A_t^r. When $t \neq T$, we have to solve

$$\min \left\|X_t - D^c A_t^c - D_t^r A_t^r\right\|_F^2 \tag{14}$$

when $t = T$, we have to solve

$$\min \left\|X_T - D^c A_T^c - D_T^u A_T^u - D_T^r A_T^r\right\|_F^2. \tag{15}$$

We can also obtain A_t^r and A_T^r by setting the derivative of (14) and (15) equal to zero, respectively. Third, we update dictionaries. Now, we are given A_t^r, A_T^r, A_T^u, D_t^r, and D_T^u, and D^c is optimized as

$$\min \|\mathcal{X} - D^c \mathcal{A}\|_F^2 \quad \text{s.t. } \left\|d_i^c\right\|_2^2 \leq 1. \tag{16}$$

$\mathcal{X} = [\mu X_1, \ldots, \mu X_{T-1}, \mu(X_1 - D_1^r A_1^r), \ldots, \mu(X_{T-1} - D_{T-1}^r A_{T-1}^r), \ldots, X_T, X_T - D_T^u A_T^u - D_T^r A_r]$, and $\mathcal{A} = [\mu A_1^c, \cdots \mu A_{T-1}^c, \mu A_1^c, \ldots, \mu A_{T-1}^c, A_T^c, A_T^c, A_T^c]$. We can get the updated D^c by solving (16) by the Lagrange dual method. Then, we fix D^c, D_T^r, A_T^c, A_T^u, and A_t^r, and D_T^u is optimized as

$$\min \left\|\mathcal{X}_T^u - D_T^u \mathcal{A}_T^u\right\|_F^2 \quad \text{s.t. } \left\|d_{T,i}^u\right\|_2^2 \leq 1. \tag{17}$$

$\mathcal{X} = [X_T, X_T - D^c A_T^c, X_T - D^c A_T^c - D_T^r A_T^r]$ and $\mathcal{A}_T^u = [A_T^u, A_T^u, A_T^u]$. Finally, we fix D^c, D_T^u, A_t^c, A_T^u, and A_t^r, and D_t^r is optimized as

$$\min \left\|\mathcal{X}_t^r - D_t^r A_t^r\right\|_F^2 \quad \text{s.t. } \left\|d_{t,i}^r\right\|_2^2 \leq 1 \tag{18}$$

when $t = T$

$$\mathcal{X}_t^r = X_T - D^c A_T^c - D_T^u A_T^u \tag{19}$$

when $t \neq T$

$$\mathcal{X}_t^r = X_t - D^c A_t^c. \tag{20}$$

Then, we get D_T^u and D_t^r by solving (17) and (18) similarly as (16).

Fig. 1. Some instances from the AID data set. (Left to Right) Image label is airport, bridge, farmland, parking, pond, river, and stadium.

V. Experiments

To evaluate the performance of HR-MTDL, we conduct experiments on the AID data set, UC-Merced data set, WHU-RS19 data set, and RSSCN7 data set, which have been popularly used for evaluating RS image recognition. We consider learning the classification model for each data set as a task. The AID data set consists of 10 000 images of 30 classes, and all the images have the same size of 600×600 pixels. The UC-Merced data set contains 2100 images of 21 classes with a fixed size of 600×600 pixels. The WHU-RS19 data set includes 950 images of 19 classes, and all the images also have the size of 600×600 pixels. The RSSCN7 data set has 2800 images of seven classes, and each image has a size of 400×400 pixels. Some samples of the AID data set are shown in Fig. 1. We divide our experiments into two parts. In the first part, we first choose one data set as the target data set and the other three as the source data sets. Then, the samples in the target data set are randomly separated into two equal-size subsets. Finally, all the samples in the source data sets and one subset of the target data set are labeled to train model while the other subset is used for testing. We use GIST feature as the image descriptor and compare our proposed method HR-MTDL with single-task dictionary learning (STDL) and LR-STDL. In the second part, we first change the rate of training samples in the target data set. Then, we get six different midlevel features by combining six types of feature coding approaches (e.g., BOVW [4], improved Fisher vector [21], locality-constrained linear coding [22], latent Dirichlet allocation [23], SPM [5], and vector of locally aggregated descriptors [24]) with local feature descriptor scale invariant feature transform. Ultimately, we conduct six pairs of contrast experiments by comparing our method with one of the mentioned approaches based on the obtained feature. For all the methods in our experiments, regularization parameters λ and η are turned from the candidate set $\{1 \times 10^i | i = -7, \ldots 7\}$. γ and μ are tuned from the set $\{1 \times 10^i | i = -10, \ldots -1\}$. The parameter k which is the number of neighbors in k-nearest neighbors for computing Hessian is set to 30 in all experiments. As for the size of dictionaries, we find that the performance of the model has a different sensibility to dictionary size in different target data sets. We consider D^c and D_t^r have the same size which is the half of D_T^u. The performance is measured by overall accuracy and confusion matrix.

Fig. 2 illustrates the average confusion matrix of two algorithms when the AID data set is the target data set. From the confusion matrix, we can see that HR-MTDL obtains higher recognition accuracy than LR-MTDL in most classes.

Fig. 3 reveals the influence of dictionary size to the mean recognition accuracy. The left and right, respectively, represent the classification performance of four methods when

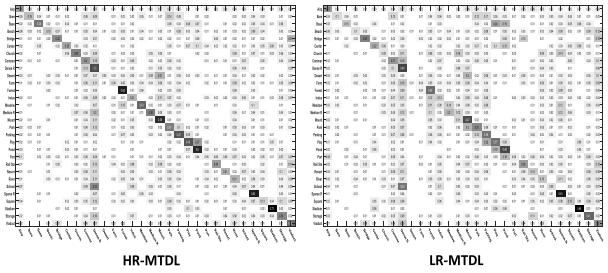

Fig. 2. Confusion matrix of HR-MTDL and LR-MTDL.

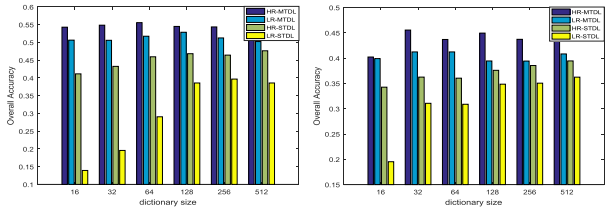

Fig. 3. Recognition accuracy versus different dictionary size.

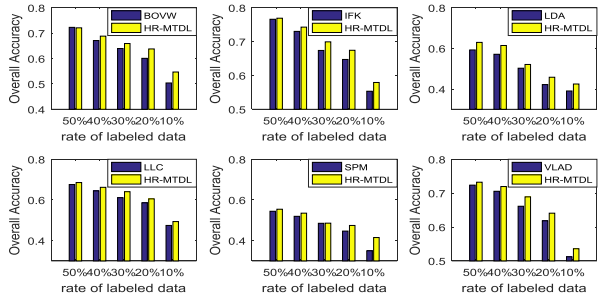

Fig. 4. Classification results versus six pairs of contrast experiments under different rates of labeled data in UC-Merced target data set.

RSSCN7 data set and WHU-RS data set are the target data sets. The horizon axis shows the dictionary size, and the vertical axis shows the mean average precision. In Fig. 3, we can see that there is an upper trend of recognition accuracy with the increase in dictionary size. When the size reaches 128, recognition accuracy seems to be stable. Overall, the proposed HR-MTDL achieves the best performance.

The recognition results of six pairs of contrast experiments are shown in Fig. 4 in which we can see that our method outperforms the compared approach in every pair of the comparative experiment. It is obvious that the recognition accuracy decreases with the reduction of training samples of the target data set, but the less the training samples, the better our method performs than the compared approach, which reflects the effectiveness of source data sets, thus verify the merits of our proposed method.

VI. Conclusion

In this letter, we incorporate Hessian regularization into multitask dictionary learning and propose the HR-MTDL. Hessian regularization has a richer null space than the graph Laplacian, thus it is superior to graph Laplacian for modeling the local geometry of compact support of the marginal distribution. Extensive classification experiments on four RS image data sets prove the effectiveness of our proposed method.

References

[1] J. A. dos Santos, O. A. B. Penatti, and R. da Silva Torres, "Evaluating the potential of texture and color descriptors for remote sensing image retrieval and classification," in *Proc. VISAPP*, 2010, pp. 203–208.

[2] M. J. Swain and D. H. Ballard, "Color indexing," *Int. J. Comput. Vis.*, vol. 7, no. 1, pp. 11–32, 1991.

[3] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[4] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, Oct. 2003, pp. 1470–1477.

[5] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2006, pp. 2169–2178.

[6] Y. Yang and S. Newsam, "Spatial pyramid co-occurrence for image classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 1465–1472.

[7] A. M. Cheriyadat, "Unsupervised feature learning for aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 439–451, Jan. 2014.

[8] Y. Luo, Y. Wen, D. Tao, J. Gui, and C. Xu, "Large margin multi-modal multi-task feature extraction for image classification," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 414–427, Jan. 2016.

[9] F. Zhang, B. Du, and L. Zhang, "Saliency-guided unsupervised feature learning for scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 2175–2184, Apr. 2015.

[10] O. A. Penatti, K. Nogueira, and J. A. dos Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2015, pp. 44–51.

[11] M. Castelluccio, G. Poggi, C. Sansone, and L. Verdoliva. (2015). "Land use classification in remote sensing images by convolutional neural networks." [Online]. Available: https://arxiv.org/abs/1508.00092

[12] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, pp. 1–13, 2018.

[13] G. Matasci, M. Volpi, M. Kanevski, L. Bruzzone, and D. Tuia, "Semi-supervised transfer component analysis for domain adaptation in remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 7, pp. 3550–3564, Jul. 2015.

[14] Y. Luo, Y. Wen, and D. Tao, "Heterogeneous multitask metric learning across multiple domains," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 9, pp. 4051–4064, Sep. 2018.

[15] Q. Wang, X. He, and X. Li, "Locality and structure regularized low rank representation for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, pp. 1–13, 2018.

[16] Y. Luo, D. Tao, C. Xu, C. Xu, H. Liu, and Y. Wen, "Multiview vector-valued manifold regularization for multilabel image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 5, pp. 709–722, May 2013.

[17] Q. Wang, J. Lin, and Y. Yuan, "Salient band selection for hyperspectral image classification via manifold ranking," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1279–1289, Jun. 2016.

[18] P. Peng *et al.*, "Unsupervised cross-dataset transfer learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1306–1315.

[19] K. I. Kim, F. Steinke, and M. Hein, "Semi-supervised regression using hessian energy with an application to semi-supervised dimensionality reduction," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 979–987.

[20] J. Eells and L. Lemaire, *Selected Topics in Harmonic Maps*, vol. 50. Providence, RI, USA: AMS, 1983.

[21] R. Negrel, D. Picard, and P.-H. Gosselin, "Evaluation of second-order visual features for land-use classification," in *Proc. 12th Int. Workshop Content-Based Multimedia Indexing (CBMI)*, Jun. 2014, pp. 1–5.

[22] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 1794–1801.

[23] Y. Zhong, Q. Zhu, and L. Zhang, "Scene classification based on the multifeature fusion probabilistic topic model for high spatial resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 11, pp. 6207–6222, Nov. 2015.

[24] H. Jegou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1704–1716, Sep. 2012.