Pattern Recognition 73 (2018) 1–14

Contents lists available at ScienceDirect

Pattern Recognition

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

Multi-modal feature fusion for geographic image annotation

Ke Li[a], Changqing Zou[b,c,*], Shuhui Bu[d], Yun Liang[c], Jian Zhang[e], Minglun Gong[f]

[a] Zhengzhou Institute of Surveying and Mapping, Longhai Road 66, Zhengzhou 450052, China
[b] College of Computer Science and Technology, Hengyang Normal Univerisity, Hengyang 421002, China
[c] Simon Fraser University, 8888 University Drive, Burnaby V5A 1S6, Canada
[d] Northwestern Polytechnical University, Youyi west road 127, Xi'an 710072, China
[e] Zhejiang International Studies University, Wensan road 140, Hangzhou 310012, China
[f] Memorial University of Newfoundland, St. John's, NL A1B 3X5 Canada

ARTICLE INFO

Article history:
Received 18 November 2016
Revised 23 April 2017
Accepted 30 June 2017
Available online 11 July 2017

Keywords:
Convolutional neural networks (CNNs)
Deep learning
Geographic image annotation
Multi-modal feature fusion

ABSTRACT

This paper presents a multi-modal feature fusion based framework to improve the geographic image annotation. To achieve effective representations of geographic images, the method leverages a low-to-high learning flow for both the deep and shallow modality features. It first extracts low-level features for each input image pixel, such as shallow modality features (SIFT, Color, and LBP) and deep modality features (CNNs). It then constructs mid-level features for each superpixel from low-level features. Finally it harvests high-level features from mid-level features by using deep belief networks (DBN). It uses a restricted Boltzmann machine (RBM) to mine deep correlations between high-level features from both shallow and deep modalities to achieve a final representation for geographic images. Comprehensive experiments show that this feature fusion based method achieves much better performances compared to traditional methods.

1. Introduction

Remote sensing technological development have led to an explosive growth in geographic images. These images are rich in the visual information that describes the Earth's surface scene containing geospatial objects such as buildings, roads, farms, forest and rivers. The automatic analysis and understanding of geographic images can enhance many applications in fields varying from environment studies to socioeconomic issues. It has attracted increasing attention, an insightful survey can be seen in [1].

Image annotation is an important component of a higher-level image understanding and semantic information extraction. Recently a number of works have shown that image representation is the key factor affecting image annotation. Image representation can be classified into two modalities. The first is shallow modality features representing the extrinsic visual properties of the image, such as scale invariant feature transform (SIFT) [2], Gabor [3], or histogram oriented gradients (HOG) [4]. The second is the deep modality feature which can represent the image's intrinsic semantic and structural representations of image. One example of this would be CNNs features [5–7].

Conventional algorithms [8–10] use only single shallow modality features to annotate images. In these algorithms, shallow feature vectors are usually extracted from the input image by using human-design descriptors such as local binary patterns (LBP) [11] and SIFT, to characterize a particular kind of information (e.g., texture, color, and shape). Generally, engineered shallow features have some advantages in classifying simple geospatial objects such as the sea, or airports [12,13]. They often have data dependency problems and have limited performances in classifying some complex geospatial objects. Seeing the two result images on the right of Fig. 1. Some shallow feature fusion based algorithms [14–17], may improve image annotation accuracy. But combinations of shallow modality features do not result in a good intrinsic semantic representation of geographic images. It is difficult to effectively improve the performance of geographic image annotation.

Recently deep-modality feature based algorithms such as Deep Convolutional Neural Networks (DCNNs) dominate the top accuracy benchmarks of various application [5,7,18,19]. These algorithms are able to generate robust generic and hierarchical deep features, and have advantages in classifying complex geographic images. However, they are not well-suited for all kinds of geographic images (see the evidence shown in the first row of Fig. 1). This may due to: 1) an invariance with regard to spatial transformation inherently limits the spatial accuracy [20], leading to a relatively weak abilities to capture the fine details required by ge-

[*] Corresponding author.
E-mail address: aaronzou1125@gmail.com (C. Zou).

http://dx.doi.org/10.1016/j.patcog.2017.06.036
0031-3203/© 2017 Published by Elsevier Ltd.

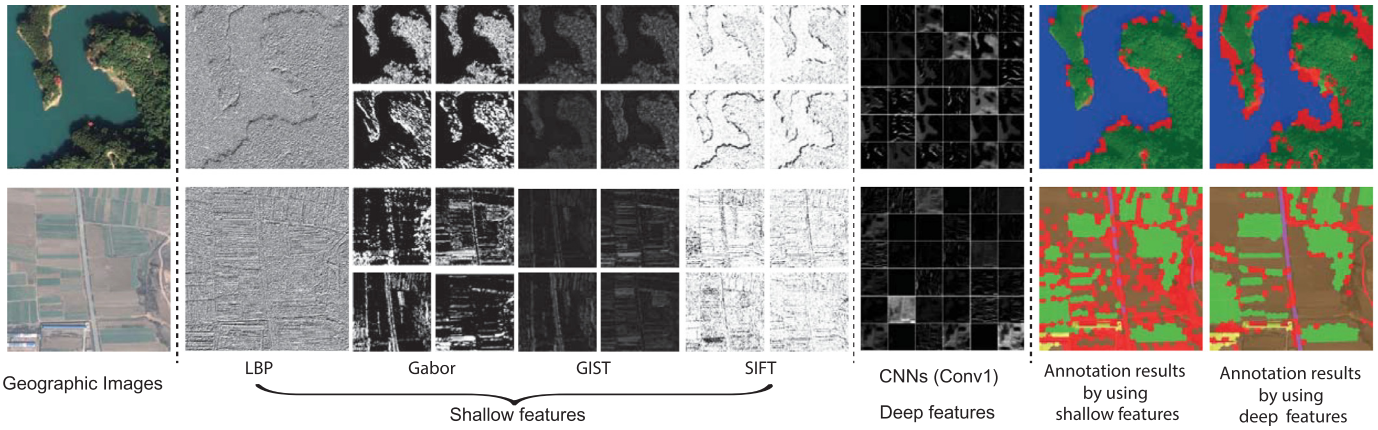

Geographic Images
LBP
Gabor
GIST
SIFT
Shallow features
CNNs (Conv1)
Deep features
Annotation results by using shallow features
Annotation results by using deep features

Fig. 1. Geographic images (left), their feature maps of shallow and deep modality (middle), and the corresponding annotation results (right); annotation errors are marked in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

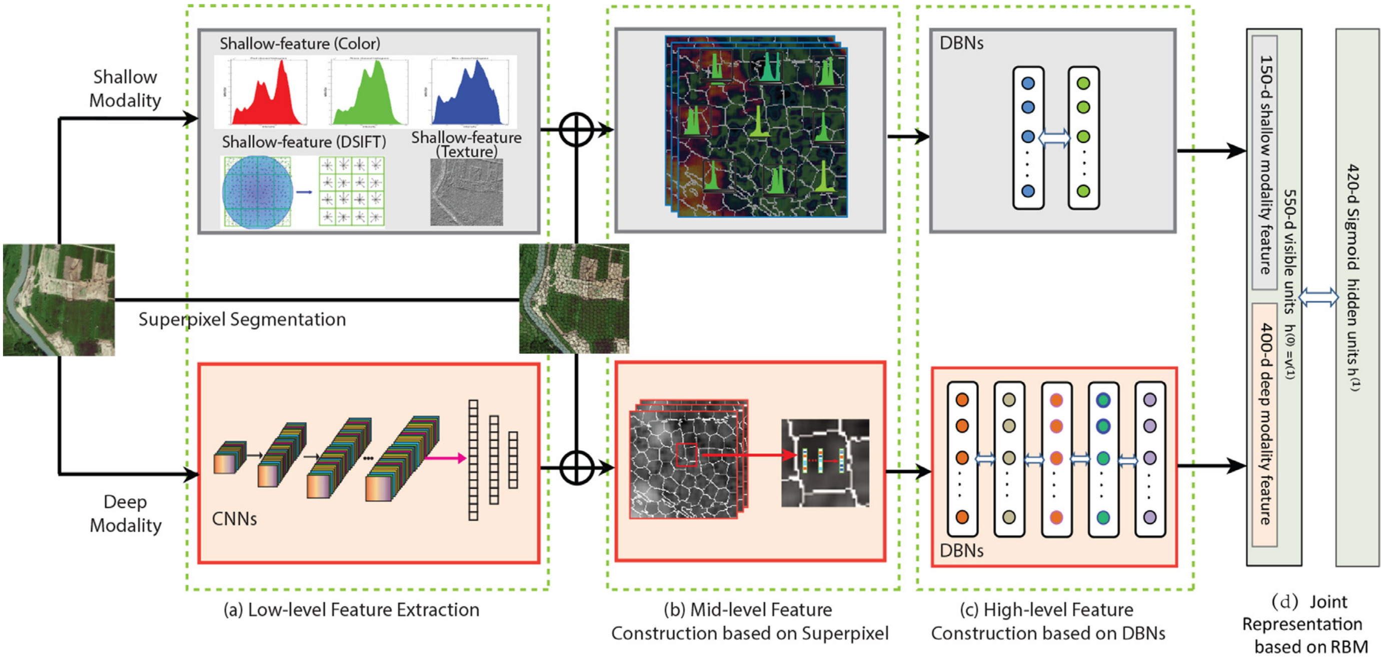

Shallow Modality
Shallow-feature (Color)
Shallow-feature (DSIFT)
Shallow-feature (Texture)
Superpixel Segmentation
Deep Modality
CNNs
DBNs
DBNs
150-d shallow modality feature
550-d visible units $h^{(0)}=v^{(1)}$
400-d deep modality feature
420-d Sigmoid hidden units $h^{(1)}$
(a) Low-level Feature Extraction
(b) Mid-level Feature Construction based on Superpixel
(c) High-level Feature Construction based on DBNs
(d) Joint Representation based on RBM

Fig. 2. Flowchart of our multi-modal feature based image representation. See the text in Section 3 for details.

ographic image annotation; and 2) most existing approaches use only single-modality features to generate image representations, which is unsuitable for geographic image data as single modality features are unable to reflect all the various characteristics of a spatial object.

The annotation task examined in this paper is to automatically label superpixels segmented from geographic input images into semantic labels, such as building, road, or river. A novel, multimodal feature fusion based framework is prosed to obtain an effective representation for each superpixel annotation. The framework consists of four sequential modules (Fig. 2): 1) a double-channel (including both shallow and deep modality) based, low-level feature extraction; 2) a mid-level feature construction within a superpixel; 3) a deep belief network (DBN) based high-level feature learning; and, 4) a restricted Boltzmann machine based (RBM) feature fusion. The four modules form a multi-modal feature evolution flow in a bottom-up structure, which combines the strengths of shallow and deep modality features. The experimental results in Section 4 show that the framework mines deep, nonlinear, correlations between deep and shallow modalities and provide a complementary enhancement for each individual modality feature.

To summarize, by analyzing the strengths and weaknesses of deep and shallow modality features, our work offers an effective solution for the problem of geographic image annotation by taking a multi-modal feature fusion. Our contributions mainly include:

(1) **Multi-modal feature construction**: as for the shallow modality features, we propose a mixed shallow feature model which combines Color, LBP, and SIFT features to represent the extrinsic visual properties of geographic images; as for the deep modality features, we design a specialized DCNN to extract the intrinsic semantic information for geographic images.
(2) **Multi-modal feature fusion**: we propose a multi-modal feature fusion model based on DBNs and RBM to build a powerful joint representation for geographic images. The model has been shown to be effective to capture both the intrinsic and extrinsic semantic information.
(3) **Open geographic image annotation benchmark dataset**: we have built a geographic image annotation benchmark

dataset named as GeographicImages60CM300, which contains 300 images (600 × 600) in six typical areas such as urban, rural, and mountain. We will make this benchmark dataset, as well as our implementation, public available to support the researches in related field.

2. Related work

A wide variety of methods have been used to investigate image annotation in recent years. Appropriate representation of geographic images is the most crucial factor for this problem. This section revisits works related to geographic image annotation. They are grouped according to the manner in which geographic images are represented.

2.1. Shallow modality feature based algorithms

Several shallow-modality feature based algorithms solutions have been proposed for the geographic image annotation problem. For instance, Luo et al. [21] performed geographic image annotation using Color and HOG. Sirmacek et al. [22] presented urban areas and buildings using the SIFT feature. Grabner et al. [23] proposed to detect vehicles from geographic images using HOG and LBP features. Wang et al. [24] employed the Gabor feature and stacked spectral features to perform hyperspectral image classification. The common feature of these methods is that they directly use one, or two, kinds of extrinsic visual properties to represent the geospatial objects, without involving a training process. These methods may obtain a good result on some kinds of simple geospatial objects. For example, Luo et al. [21] achieve a very good annotation accuracy of the “sea” in geographic images. They are usually not powerful for a broad range of geospatial objects.

2.2. Deep modality feature based algorithms

Algorithms utilizing DCNNs to represent geographic images have been used for annotation. Farabet et al. [25] propose using a multi-scale convolutional network combined with a conditional random fields (CRF) model to parse the images. Yue et al. [26] used the algorithm to preform hyperspectral image classification by using a logistic regression classifier on spectral and spatial feature maps. It obtains the feature maps using a classic CNN. The methods in [27,28] are to detect vehicles from high-resolution images by using hybrid CNNs frameworks which are capable of extracting multi-scale features. Generally, deep-modality feature based algorithms have better annotation accuracy compared to shallow-feature based ones on a larger range of geospatial objects. This is particularly true for complex geographic scenes, or objects, such as buildings. Their performances on some simple geographic scenes, on the other hand, are less than ideal (See the Fig. 12(c)).

2.3. Feature fusion based algorithms

There are also some other frameworks which combine/fuse several features to improve the performance of geographic image annotation. For example, Zhang et al. [16] and Tuia et al. [17] concatenated multiple features by employing a vector-stacking (VS) strategy to provide the data required by the classifier for geospatial objects. VS is simple to execute and has a potential to enhance the discrimination between similar geospatial objects [29]. It does not mine deep correlations of various features very well. To overcome this limitation, research in [30,31] proposed using a framework based on a combination of deep Boltzmann machines (DBM) and RBM to learn an image representation over multiple modality features. It first employs DBM to learn multi-modal features from unlabeled data and then uses RBM to find a common space representation for different input modalities. In addition to the DBM+RBM framework, joint spare coding [32,33] and auto-encoder [34] are also popular tools for multi-modal feature fusion.

Multi-view learning methods [35,36] are used to fuse different kinds of features in application scenarios including image classification and face recognition. The key idea of multi-view learning methods and multi-modal fusion methods is to embed inputs from different domains into a new latent common space, which can then better mine non-linear correlations of different representations.

2.4. Geographic image annotation based algorithms

Geographic image annotation is usually carried out in feature space. Effective feature representation is very important to construct high-performance image annotation systems. Recently considerable efforts have been made to develop various feature representations to annotate different types of objects in satellite and aerial images such as color, Haar, SIFT, LBP, or HOG. Porway et al. [37] combined color and edge features with object-level features in a hierarchical contextual model for geospatial image annotation. Markususe et al. [38] applied AdaBoost classifications based on Haar, and Textons features for semantic labelling on the ISPRS benchmark. Cheriyadat et al. [39] used SIFT feature and graph sparse coding algorithm to annotate geospatial objects in aerial images. Kembhavi et al. [40] used multi-scale HOG features computed to annotate vehicles in San Francisco images from Google Earth, and showed HOG to outperform SIFT in complex city environments. Grabner et al. [23] used boosting methods based on LBP and HOG to detect vehicles.

When only a small training set is available, using engineered shallow features and traditional classifiers is a reasonable approach. But if there are large numbers of samples for each class, learning the deep features from the training samples is more advisable. Deep Learning is currently fashionable for automatically learning robust features from the raw data. Chen et al. [41] relied on stacked autoencoders, trained to reconstruct PCA-compressed hyperspectral signals. The network is then fine-tuned by backpropagating errors from a softmax loss on top of the stacked autoencoders. Both Castelluccio et al. [42] and Marmanis et al. [43] fine-tuned pre-trained CNNs to annotate geospatial images. Paisitkriangkrai et al. [44] proposed a system based on CNNs trained on the Vaihingen challenge data set to perform semantic labeling. CNN potential is clearly shown by combining the features extracted from the CNN with random forest classifiers, standard appearance descriptors, and conditional random fields, performing structured prediction on the probabilities given by the classifier. Sherrah et al. [45] using a deep FCN with no downsampling to annotate high-resolution aerial imagery on a Vaihingen challenge data set, eliminating the need for either deconvolution, or interpolation.

Although many techniques have performed well in geographic image annotation, there is yet room for improvement. That is: 1) most prior algorithms only leverage partial information about geographic images, either shallow features or deep features, for annotation; and 2) most existing multi-modal fusion models are ineffective in discovering highly nonlinear relationships between features across different modalities due to their relatively simple model structures.

The proposed framework is different from present methods as it utilizes both shallow and deep modality features to annotate. This framework has the advantages of both the shallow and deep modality based algorithms. Specifically, the shallow modality feature channel in our framework combines SIFT, Color and LBP features which encode a complete extrinsic semantic information (including local invariant, color, and texture information) of an image; the deep modality feature channel use a powerful DCNNs to capture abundant intrinsic semantic information. The two-channel (extrinsic and intrinsic) semantic information provides sufficient discrimination power for ensuring good annotation results. Compared to previous fusion model, feature fusion is addressed by combining RBM and DBNs to deeply fuse the dual-channel semantic information. RBM which learns feature representation in an unsupervised manner and has demonstrated to be promising in building high-level feature descriptors is used to discovers the nonlinear correlations between the deep, and shallow, modality features. More importantly, noise interference such as weather, illumination intensity, and building shadows usually result in many missing values in geospatial images. Probabilistic model based RBM can handle the problem of missing values and generate samples in a natural way [36].

3. Multi-modal feature based image representation and image annotation

As illustrated in Fig. 2, for an input image I, we first use the linear iterative clustering (SLIC) algorithm [46] to segment I into a set of superpixels $\mathcal{S}$. For each superpixel $S_i \in \mathcal{S}$, we utilize a shallow modality channel and a deep modality channel to respectively extract shallow features $\mathbf{V}_S$ and deep features $\mathbf{V}_D$. We then employ a RBM model to generate the final representation of S_i by fusing $\mathbf{V}_S$ and $\mathbf{V}_D$. In our framework both shallow features and deep features are achieved by a feature evolution process which is formed by three sequential modules for low-level, mid-level, and high-level feature extraction, respectively. The shallow features $\mathbf{V}_S$ encode the extrinsic visual properties of a geographic image, whereas the deep features $\mathbf{V}_D$ encode the intrinsic semantic information. The two mutually complementary sets of features are fused together as the final representation $\mathbf{V}_J$ of S_i. We next describe individual modules of the proposed framework.

3.1. Shallow modality feature

The generation of $\mathbf{V}_S$ contains three sequential steps. We first extract multiple types of low-level features including the LBP, SIFT, and Color features for each pixel of the input image I; and then we generate a mid-level feature vector $\mathbf{V}_S^m$ for each super-pixel S_i by integrating the low-level features of all the pixels within S_i; lastly we use DBNs model to further construct a high level feature vector $\mathbf{V}_S$ from $\mathbf{V}_S^m$ for each super-pixel S_i.

3.1.1. Low-level shallow feature

SIFT Feature: SIFT is an image descriptor for image-based matching and recognition developed by Lowe [2], which is widely used for various purposes in computer vision related to point matching between different views of a 3-D scene and view-based object recognition. The original formulation of SIFT comprised a method for detecting interest points from a grey-level image at which statistics of local gradient directions of image intensities were accumulated to give a summarizing description of the local image structures in a local neighborhood around each interest point. The extended dense SIFT is a descriptor applied at dense grids rather than detected interest points. Dense SIFT has been shown to lead to better performance for various tasks such as object categorization, texture classification, image alignment and biometrics. In this paper, we use the dense SIFT implemented in [47] to extract the SIFT feature of each pixel, i.e., the SIFT feature of a pixel p is extracted using p as the key-point. In our implementation, the width in pixels of a spatial bin is set to 2 and we obtain a 128 dimensional feature vector $\mathbf{f}_p^s$ for a given pixel p.

LBP Feature: The LBP operator [48] has been widely used in various applications. It has been proven to be highly discriminative and has the advantages of invariance w.r.t. monotonic gray-level changes and computational efficiency. LBP has been found improving the detection performance considerably when it is combined with some other local image gradient based descriptors [49]. This motivated us to combine LBP with the SIFT descriptor to construct the low-level local feature around points of interests. Similar with [48], we compare each pixel of the input image to each of its 8 neighbors along a clockwise circle and generate a 8-digit binary number, and then use this 8-digit binary number (corresponding to a decimal number within 0–255) as the low-level feature of each pixel.

Color Feature: Color feature is an important feature in geographic images. Both SIFT and LBP features are extracted from gray images and do not encode any color information. Color features can be a strong supplement to represent some typical objects related to some special colors in geographic images, e.g., blue sea or green forest. In our method, we obtain the low-level color feature on each pixel p as the following steps: 1) convert the RGB space of the input image I to the $l\alpha\beta$ space; 2) normalize the color value of p in each channel of $l\alpha\beta$ between [0, 1]; and 3) use the normalized color values as the three demesnial low-level color feature of the pixel.

3.1.2. Mid-level shallow feature

All the above three types of low-level features are defined at pixel-level, which describe either a single pixel or a relatively small neighborhood surrounding a center pixel. They often cannot provide sufficient semantic characteristics required by the annotation task. Therefore, it is necessary to construct higher level features from the low-level features of individual pixels to represent the segmented superpixels. Our mid-level feature descriptor is designed to fulfill this objective, which characterizes the local image content within each super-pixel. For a super-pixel S_i in the input image I, we generate its mid-level feature vector $\mathbf{V}_S^m$ by concatenating three feature vectors $\mathbf{F}_{S_i}^s$, $\mathbf{F}_{S_i}^l$, and $\mathbf{F}_{S_i}^c$, i.e., the mid-level shallow feature vector of S_i has the following form:

$$\mathbf{V}_S^m = [\mathbf{F}_{S_i}^s, \mathbf{F}_{S_i}^l, \mathbf{F}_{S_i}^c], \tag{1}$$

where $\mathbf{F}_{S_i}^s$, $\mathbf{F}_{S_i}^l$, and $\mathbf{F}_{S_i}^c$ are respectively extracted from the low-level SIFT, LBP, and color features of all the pixels within S_i.

In Eq. (1), $\mathbf{F}_{S_i}^s$ is computed by the L_2 normalization of the average of the SIFT feature vectors of all the N_{S_i} pixels within S_i, having the following formulation:

$$\mathbf{F}_{S_i}^s = \| \sum_{p \in S_i} \mathbf{f}_p^s / N_{S_i} \|_2. \tag{2}$$

$\mathbf{F}_{S_i}^s$ has the same number of dimensions as the low-level SIFT vector $\mathbf{f}_p^s$ of pixel p (128D). The second feature component $\mathbf{F}_{S_i}^l$ is obtained by computing the histogram, over the super-pixel S_i, of the frequency of each number (8-digit binary number) in the low-level LBP features occurring. The initial vector of $\mathbf{F}_{S_i}^l$ created from the statistic histogram has 256 dimensions since there are 256 types of 8-digit binary numbers in total in the LBP features. In our implementation, after obtaining the 256 dimensional initial vector, we adopt the principal components analysis algorithm (PCA) [50] to reduce the initial 256 dimensional-vector to a more compact 80-dimensional feature vector. The final vector $\mathbf{F}_{S_i}^l$ is the L_2 normalization of the 80 dimensional feature vector. The last feature component $\mathbf{F}_{S_i}^c$ is related to the statistic histogram, in terms of color channel, of the low-level color features of all the pixels within S_i. Specifically, for each channel of all the color features, we quantize the normalized values into 25 bins. We generate a 75 bin histogram by concatenating the histograms from three channels. The L_2 normalization of this 75-dimensional vector forms the vector $\mathbf{F}_{S_i}^c$.

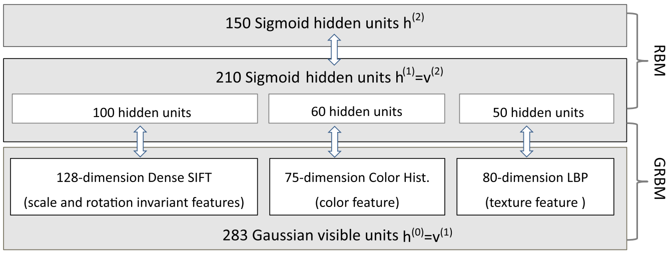


Fig. 3. DBNs architecture for the high-level shallow features.

3.1.3. High-level shallow feature

Many works on DBNs [51,52] have shown that it is possible to capture high-level non-linear features via multiple-layer network without labeled data. In order to express extrinsic visual information of geospatial image sufficiently, we therefore employ the DBNs model to further construct high-level shallow features from the mid-level shallow features.

Our DBNs is a deep belief network of three layers with architecture shown in Fig. 3. The bottom layer is the input layer with Gaussian visible units and connects to the second layer of Sigmoid hidden units. These two layers are composed of Gaussian restricted Boltzmann machines (GRBMs) and restricted Boltzmann machines (RBMs). The two models [53,54] are defined as:

$$GRBM(\mathbf{v},\mathbf{h};\theta) = \sum_{i=1}^{D}\frac{(v_i-a_i)^2}{2\sigma_i^2} - \sum_{i=1}^{D}\sum_{j=1}^{F}\frac{v_i}{\sigma_i}w_{ij}h_j - \sum_{j=1}^{F}b_jh_j. \quad (3)$$

$$RBM(\mathbf{v},\mathbf{h};\theta) = -\sum_{i=1}^{D}\sum_{j=1}^{F}v_iw_{ij}h_j - \sum_{i=1}^{D}a_iv_i - \sum_{i=1}^{D}b_jh_i \quad (4)$$

where $\mathbf{v}$ is visible units ($\mathbf{v}\in\mathbb{R}^D$ in GRBM and $\mathbf{v}\in\{0,1\}^D$ in RBM), σ_i is adopted to model the variation of visible unit i, which is obtained through analyzing the distribution of the input data, and $\mathbf{h}\in\{0,1\}^F$ is stochastic hidden units, with each visible unit connected to each hidden unit. $\theta=\{\mathbf{a},\mathbf{b},\mathbf{W}\}$ are the model parameters. w_{ij} represents the symmetric interaction between visible unit v_i and hidden unit h_j, b_i and a_j are the biases, and D and F are the numbers of visible and hidden units.

We propose to train a separate GRBM for each of the three types of middle-level features (i.e., $\mathbf{F}_{S_i}^c$, $\mathbf{F}_{S_i}^l$, and $\mathbf{F}_{S_i}^s$). This enables us to reduce not only the number of free parameters but also the computational cost for training. The output of the GRBMs are concatenated to form the input to the next layer, which is a single RBM with Sigmoid units in both layers (210 inputs and 150 outputs for the case in Fig 3). The output of the last layer is used as the high-level shallow feature $\mathbf{V}_S$.

We pre-train the DBNs model greedily in a layer-by-layer manner. In our work, GRBM or RBM is trained by contrastive divergence (CD) [55], the detailed algorithm process is described in Algorithm 1. The core part is to compute the gradient of likelihood by 1-step contrastive divergence (CD-1), which uses Gibbs sampling to approximate the intractable true gradient. Specifically, when given a number of samples for the input layer $\mathbf{v}_+$, we simulate the model for 1.5 cycles and collect activation of visible and hidden units $\mathbf{h}_+$, $\mathbf{h}_{\_}$, and $\mathbf{v}_{\_}$ for the first and last upward half-cycle.

In our implementation, we set the bias on the hidden units to zero, and the batch size $Size_{batch}$ is 100. To accelerate learning, we add momentum to the gradient with $\alpha = 0.5$. The global learning rate η is 0.1 for Gaussian-RBMs and a much smaller rate of $\eta = 0.05$ for the top-level RBM. The training generally converges in $T = 1000$ epochs.

Algorithm 1 Pre-train RBM model.

```
Input: Sample a minibatch of $M$ examples from the training set $v^{(1)}, ..., v^{(M)}$;
Output: model parameters $\Theta = \{\theta_1, \theta_2, \ldots, \theta_K\}$;
  Parameters: learning rate $\eta$, momentum $\alpha$;
  initialize $\Theta$ (see text for details);
  for $t = 1, \ldots, T$ do
    for $\mathbf{v}_+ = v^{(1)}, ..., v^{(M)}$ do
      compute hidden unit $\mathbf{h}_+ = sigmoid(\mathbf{W}\mathbf{v}_+ + \mathbf{a})$;
      compute visible unit $\mathbf{v}_{\_} = sigmoid(\mathbf{W}^\top\mathbf{h}_+ + \mathbf{b})$;
      compute hidden unit $\mathbf{h}_{\_} = sigmoid(\ \mathbf{W}\mathbf{v}_{\_} + \mathbf{a})$;
      compute CD-1 gradients $\Delta\theta^{(t)}$ :

      $\Delta\mathbf{W}$ = CD ($\mathbf{v}^\top\mathbf{h}$), $\Delta\mathbf{a}$ = CD ($\mathbf{h}$), $\Delta\mathbf{b}$ = CD ($\mathbf{v}$);

      add momentum $\Delta\hat{\theta}_k^{(t)} = \Delta\theta_k^{(t)} + \alpha\Delta\theta_k^{(t-1)}$
      adjust model $\theta_k = \theta_k + \eta/Size_{batch}\Delta\hat{\theta}_k^{(t)}, k = 1..K$
    end for
  end for
```

3.2. Deep modality feature

The generation of the deep modality feature $\mathbf{V}_D$ contains three sequential phrases similar to the shallow modality feature $\mathbf{V}_S$. Fig. 4 gives a detailed illustration on the construction of the deep modality feature. That is, the process starts from a classic convolutional neural network where low-level hierarchies of deep feature (maps) are extracted across various layers of the network. In the succeeding phase, sets of selected deep feature maps, after up scaled to the same size as the input image I, are accumulated and statistics are further made to generate a middle-level summary of the deep features within the spatial ranges corresponding to each superpixel S_i. In the last phase, DBNs is employed to extract more concentrated and representative high-level deep features from the middle-level deep features.

3.2.1. Low-level hierarchical deep feature

Good deep modality features can represent intrinsic hierarchical information of image [25]. To achieve a rich hierarchical semantic representation of a geographic image, we employ the CNNs to perform the low-level deep modality feature extraction. In a typical deep CNN network for image presentation, the network is usually trained with multiple stages. The input and output of each stage are sets of arrays called feature maps. The output feature map is treated as a further abstraction of the input feature map. As shown in Fig. 4 (a), each stage often contains four parts: convolutional operation (* operator), non-linearity transformation such as *sigmoid* function or *ReLU* function, local response normalization (*LRN* operator) and feature pooling (*pool* operator).

In order to balance the quality of the features and the efficiency, we only select the outputs from partial layers to produce the geospatial image descriptors. In our work, the low-level hierarchical deep feature consists of the feature maps of Pool2, Conv4, and Pool5 layers. Let F_i, P_j be the feature maps of convolutional and pool layers, separately, where $i\in\{1,\ldots,5\}$ and $j\in\{1, 2, 5\}$, so the feature maps of Pool2, Conv4, and Pool5 layers can be obtained by:

$$P_2 = pool(LRN(ReLU(W_2 * P_1 + b_2))) \quad (5)$$

$$F_4 = ReLU(W_4 * F_3 + b_4) \quad (6)$$

$$P_5 = pool(ReLU(W_5 * F_4 + b_5)) \quad (7)$$

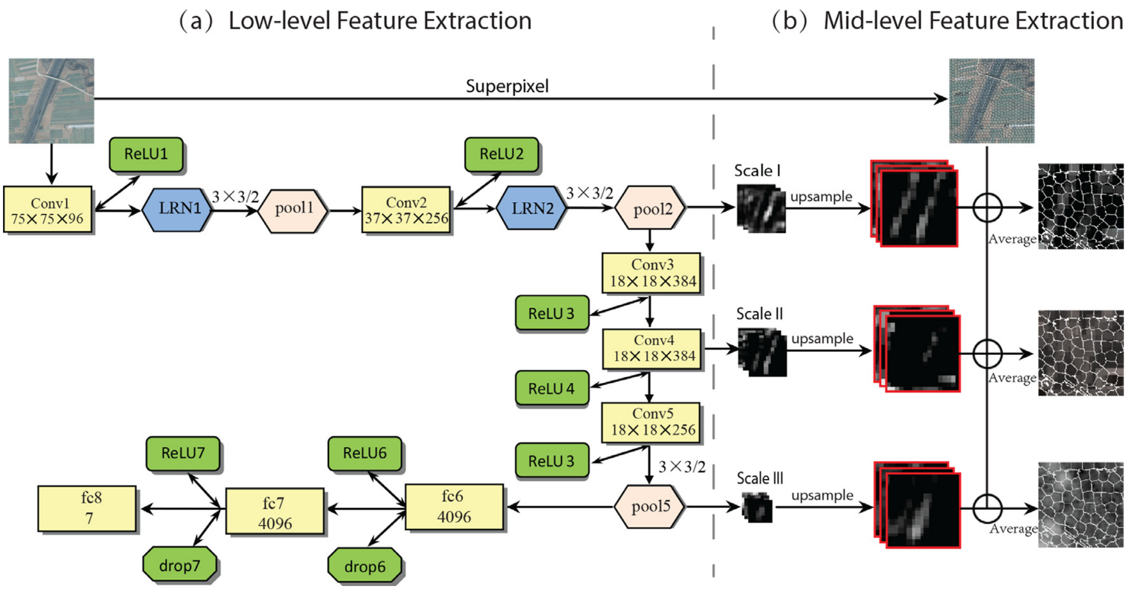


Fig. 4. Our CNN architecture for deep modality feature learning.

where b_l is the bias parameter, W_l is the convolutional kernel, $l \in \{2, 4, 5\}$.

We pre-train the CNNs on our geospatial image dataset using image-level annotations only. Our CNNs architecture is similar to the Alex network [19], However, to adapt our CNNs to the geospatial object annotation, we adjust the parameters of CNNs. The details of architecture are introduced as follows:

The input geospatial image size is 300×300, and we continue stochastic gradient descent (SGD) training of the CNN parameters. The mini-batch size is 30, and the initial learning rate is 0.0006. We train 45K iterations and replace the ImageNet specific 1000-way classification layer with a randomly initialized 7-way classification layer (where 6 is the number of geospatial object classes, plus 1 for background).

As shown in Fig. 4, there are five convolutional layers, the parameter configuration of each convolutional layer can be described concisely by layer notations with layer sizes: Conv1 (patch size/stride: $5 \times 5/4$; Feature Map: $75 \times 75 \times 96$); Conv2 (patch size/stride: $5 \times 5/1$; Feature Map: $37 \times 37 \times 256$); Conv3 (patch size/stride: $3 \times 3/1$; Feature Map: $18 \times 18 \times 384$); Conv4 (patch size/stride: $3 \times 3/1$; Feature Map: $18 \times 18 \times 384$); Conv5 (patch size/stride: $3 \times 3/1$; Feature Map: $18 \times 18 \times 256$). The max-pooling layers [19,56] follow the first, second and fifth convolutional layers, which summarize the activities of local patches of neurons in convolutional layers. All the pooling layers summarize a 3×3 neighborhood and use a stride of 2.

3.2.2. Mid-level hierarchical deep feature construction

We utilize two steps to generate a superpixel-level hierarchical deep feature descriptor from the middle level shallow features, which is shown in Fig. 4(b). First, after all the feature maps of all layers in the low level deep feature extractor CNNs are generated, we upscale the feature maps of Pool2, Conv4, and Pool5 layers to the same size as input image and then concatenate them to produce a three dimensional arrays $O \in \mathbb{R}^{H \times W \times N}$, where the three dimensions are respectively the height of images H, the width of images W, and the number of feature maps N.

Therefore, for a pixel p in the input image, we obtain a representative feature vector denoted as $O_p \in \mathbb{R}^N$ ($N = 896$ in our work). In the second step, for a super-pixel S_i in the input image I, we compute the mid-level feature vector $\mathbf{V}_D^m$ by the L_2 normalization of the average of the feature vectors O_p of all the N_{S_i} pixels within S_i, having the following formulation:

$$\mathbf{V}_D^m = \left\| \sum_{p \in S_i} O_p / N_{S_i} \right\|_2 \tag{8}$$

where $\mathbf{V}_D^m$ has the same dimension length as the hierarchical deep feature vector O_p of pixel p.

3.2.3. High-level hierarchical deep feature construction

Similar as the shallow modality channel, we still employ the DBNs to construct high-level deep features from the middle-level deep features. The architecture of our DBNs is illustrated as Fig. 5, which is a six-layer model. The bottom most two layers compose a Gaussian RBM which encodes the three kinds of middle-level deep features with the hidden layer activities. The output of these Gaussian RBMs are concatenated as the input to the next layer, which is just a single RBM with Sigmoid units in both layers. The output of this single RBM is further used as the input of three stacked RBMs which are used to boost the discrimination ability as well as reduce the redundant information of hierarchical deep features. We pre-train the parameters of the DBNs using the similar steps as Section 3.1.3. The final extracted higher-level features $\mathbf{V}_D$ are outputted from the last layer and have 400 dimensions.

3.3. Multi-modal feature fusion and annotation

We fuse the two modality high level features by a RBM model with a single layer network with architecture shown in Fig. 2(d). The input of RBM model is a 550-dimensional feature vector concentrating $\mathbf{V}_S$ and $\mathbf{V}_D$, the joint layer contains 420 hidden units. The parameters of the model are also pre-trained using the similar steps as Algorithm 1. Our multi-modal feature fusion model finally produces a 420-dimensional feature $\mathbf{V}_J$.

After obtaining $\mathbf{V}_J$, we perform one-versus-all annotation by using softmax regression. For any superpixel S_i, let $\hat{P}_{S_i}$ be the normalized prediction vector, and then we compute normalized predicted probability distributions $\hat{P}_{S_i,a}$ of class a by using the softmax function. More specifically,

$$\hat{P}_{S_i,a} = \frac{e^{\mathbf{W}_a^T \mathbf{V}_J}}{\sum_{i \in classes} e^{\mathbf{W}_i^T \cdot \mathbf{V}_J}} \tag{9}$$

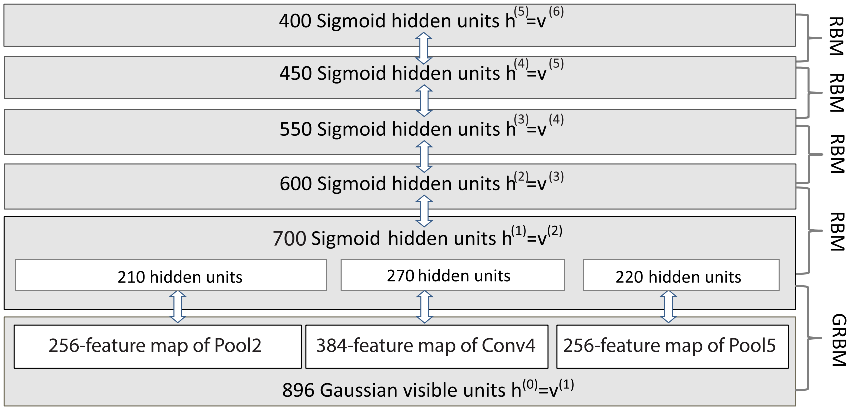

400 Sigmoid hidden units $h^{(5)}=v^{(6)}$
450 Sigmoid hidden units $h^{(4)}=v^{(5)}$
550 Sigmoid hidden units $h^{(3)}=v^{(4)}$
600 Sigmoid hidden units $h^{(2)}=v^{(3)}$
700 Sigmoid hidden units $h^{(1)}=v^{(2)}$
210 hidden units
270 hidden units
220 hidden units
256-feature map of Pool2
384-feature map of Conv4
256-feature map of Pool5
896 Gaussian visible units $h^{(0)}=v^{(1)}$
RBM
RBM
RBM
RBM
GRBM

Fig. 5. DBNs architecture for the high-level deep features.

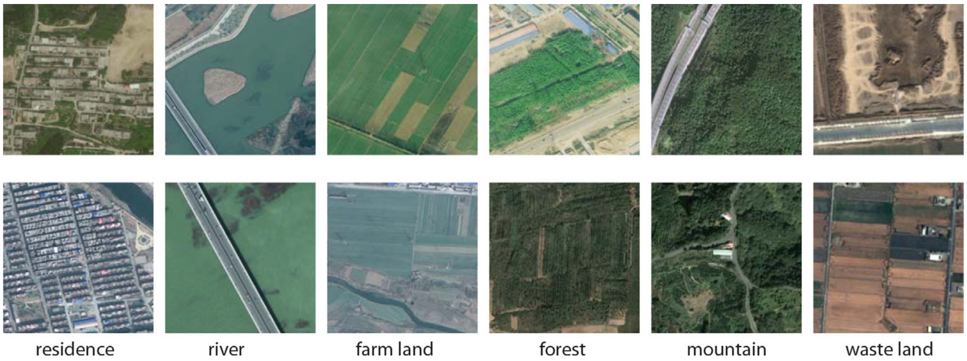


Fig. 6. Examples of the geographic images in our dataset. There are six typical labels (areas) in total for the pixels of the geographic images in the dataset.

where **W** is a temporary weight matrix used to learn the features. For each superpixel S_i, the final annotation label l_{S_i} is given by

$$l_{S_i} = \arg\max_{a \in classes} \hat{P}_{S_i,a}. \tag{10}$$

4. Experiments

Two data-sets and several experiments are conducted to evaluate the performance of the presented framework in comparison with baseline methods, which use different single-modality features. The parameters of each method were tuned for best performances. The evaluation metrics used include Average Precision (AP) for image annotation, Standard Deviation (std), and Wilcoxon Rank Sum test [57]. A deep learning toolbox[1] was implemented in which all matrix operations were carried out on a GPU with cudamat library. All experiments were conducted on a computer with 16 GB memory and Intel i7 processor.

4.1. Geographic image datasets

GeographicImages60CM300. This is a fabricated dataset that contains 300 geographic images collected from Google Map with 60-cm resolution. Each image has a resolution of 600 × 600 pixels. Each superpixel level geo-spatial object for all the collected geographic images was labeled. Fig. 7 describes the geographic image annotation procedure. Each geographic image was over-segmented using a SLIC algorithm [46]. To generate ground truth labeling data, the outline of each object is traced manually using an in-house developed interactive tool, based on which the label for each superpixel was inferred automatically. As shown in Fig. 6, six semantic labels are used, which are urban residential, rural residential, riverine, farm land, waste land, forest, and mountain.

ISPRS benchmark. The second dataset used ISPRS 2D Semantic Labeling-Vaihingen dataset [58]. The Vaihingen dataset contains 33 very high-resolution true ortho photo (TOP) tiles, varying in size from 1388 × 2555 to 3816 × 2550, and corresponding digital surface models (DSMs) derived from dense image matching techniques. Pixel-level semantic labels, for the 6 categories: “impervious surface”, “building”, “low vegetation”, “tree”, “car”, and “clutter”, are available for 16 images. The labels of the remaining images serve as private testing set for the contest. To date, April, 2017, more than 60 algorithms have reported their results.

The experiments presented below first use the GeographicImages60CM300 dataset to tune and evaluate the feature-extraction architecture. The performance comparisons are then conducted on both GeographicImages60CM300 and ISPRS 2D Semantic Labeling-Vaihingen datasets.

4.2. Architecture evaluation

4.2.1. Shallow feature combination

This set of experiments were designed to get the best combination of shallow features. Five feature types containing SIFT, GIST, Color, LBP, and Gabor from geographic images which express color, texture, and edge information in images were extracted. Ten sets of experiments using different feature combinations of the five shallow features were performed. As shown in Fig. 8(a), the image representation used the architecture named MSMF. It is a substructure of the framework shown in Fig. 2.

The feature combination results are plotted in Fig. 9, which show that, among all the two-feature combinations, Color and LBP annotation accuracies are the two best which suggests that Color

[1] http://www.adv-ci.com/blog/source/deepnet-cuda/.

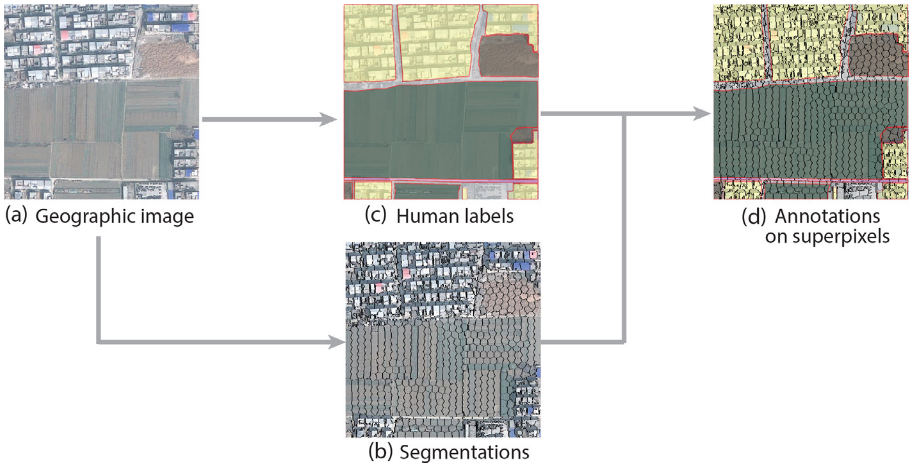


Fig. 7. The procedure for ground-truth data generation.

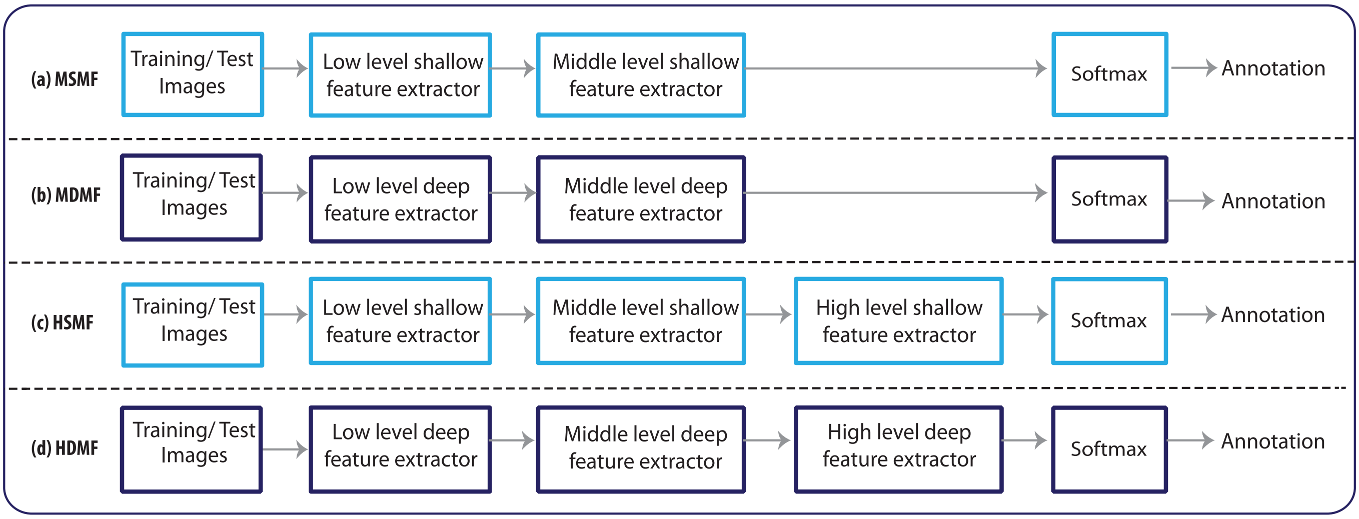

(a) MSMF
Training/ Test Images
Low level shallow feature extractor
Middle level shallow feature extractor
Softmax
Annotation
(b) MDMF
Training/ Test Images
Low level deep feature extractor
Middle level deep feature extractor
Softmax
Annotation
(c) HSMF
Training/ Test Images
Low level shallow feature extractor
Middle level shallow feature extractor
High level shallow feature extractor
Softmax
Annotation
(d) HDMF
Training/ Test Images
Low level deep feature extractor
Middle level deep feature extractor
High level deep feature extractor
Softmax
Annotation

Fig. 8. Different architectures for image representation. Architectures (a–d) are four different sub-structures of the architecture in Fig. 2.

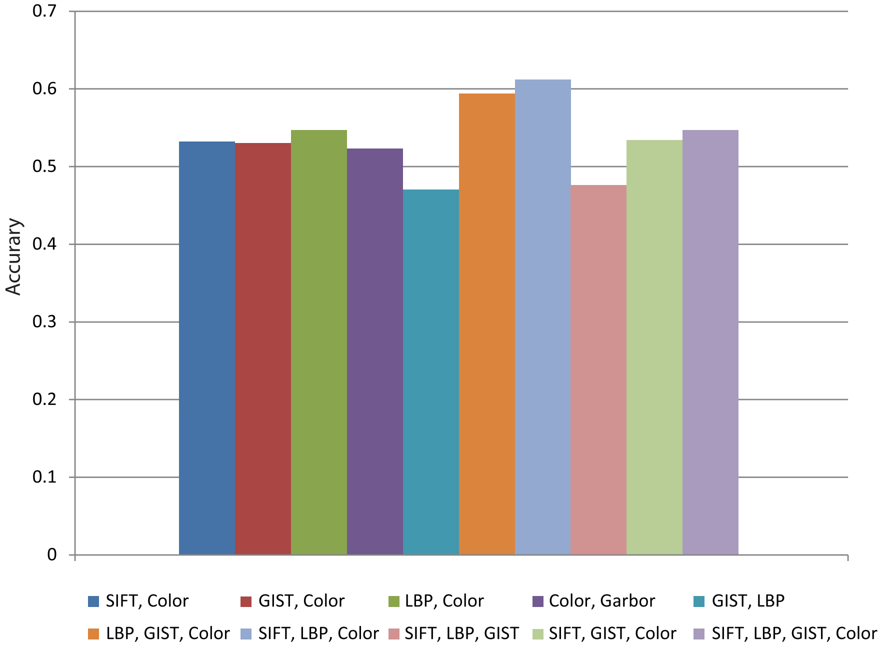


Fig. 9. Annotation accuracies of various combinations of shallow features.

Table 1
Annotation accuracies by different CNN architectures. Bold values indicate maximum performance.

Feature representations	Mean acc.	Std acc.
Conv1, 2, 3, 4, 5	74.464%	± 2.28%
Conv3, 4, 5	74.663%	±2.77%
Pooling1, 2, 5	75.519%	±2.28%
Conv4, 5; Pooling1, 2, 5	75.237%	±3.03%
Conv4, 5; Pooling2, 5	75.099%	±2.29%
Conv4; Pooling1, 2, 5	76.102%	±2.44%
Conv4; Pooling2, 5	**77.257%**	±2.58%

Table 2
Annotation accuracies by different architecture for feature extraction.

Architecture type	Mean acc.	Std acc.
MSMF	61.17%	± 2.86%
HSMF	62.864%	± 2.79%
MDMF	77.257%	± 2.58%
HDMF	79.233%	± 2.62%
MFF	**82.019%**	± 2.44%

Table 3
Annotation accuracies by different high-level modality feature representations.

	HSMF	HDMF	MFF
Building	51.7% ± 4.04%	69.2% ± 2.39%	**78.8%** ± 1.52%
River	81.5% ± 2.39%	**89.6%** ± 2.36%	88.3% ± 2.17%
Road	59.5% ± 1.08%	**83.8%** ± 2.77%	82.7% ± 2.72%
Waste Land	56.5% ± 3.19%	82.5% ±2.34%	**82.8%** ± 1.91%
Fram Land	68.7% ± 2.03%	72.0% ±3.15%	**79.1%** ± 3.09%
Forest	59.3% ± 3.98%	78.3% ±2.65%	**80.4%** ± 3.26%
Mean Acc.(Std Acc)	62.9% ± 2.79%	79.2% ± 2.62%	**82.0%** ± 2.44%

Table 4
Wilcoxon test for image annotation experiment. A positive value indicates the fusion feature outperforms a single modality feature. Negative values indicate that fusion feature does not perform significantly better.

	MFF					
	Building	River	Road	Waste Land	Farm Land	Forest
HSMF	**1**	−1	**1**	**1**	**1**	**1**
HDMF	**1**	−1	−1	−1	**1**	−1

and LBP are complementary. Another interesting finding is that the combination of SIFT, LBP, and Color features, rather than the one containing all the four types of features, achieved the best performance. Hence, this three-feature combination is used in our algorithm as the architecture low level shallow feature extractor.

4.2.2. Deep feature combination

This set of experiments were designed to get the best CNN architecture for deep features. CNNs models in Fig. 4(a) using different network settings were independently evaluated. Eight-layer feature maps including Conv1, 2, 3, 4, 5, and Pooling1, 2, 5 layers for each superpiexl segmentation were extracted. The experiments using six types of feature map combinations. The image representation used the architecture named MDMF Fig 8(b).

Four thousand 300 × 300 image were used to train the CNNN models. These training images were generated in two ways. In the first method, one part (2000 300 × 300) came from Google maps. In the second method was extracted from the geographic image dataset (about 10 overlapped sub-images were extracted from each geographic image). A classifier was built by connecting a softmax regression to the CNNs models to perform the annotations under different CNNs architectures. In each evaluation, the classifier was trained on a random sampling of 85% samples for each category of the dataset and was tested on the rest of the samples. The results in Table 1 summarize the mean accuracies for all the evaluations.

Results in previous subsection (Fig. 9) show that most shallow feature combination schemes have an accuracy of less than 60%. In comparison, the accuracy achieved by the deep feature combinations has about 20% improvement; see Table 1. The accuracies of various combinations of deep features are close and around 75%. The feature combination of Conv4, Pooling2 and 5 layers achieved the best annotation accuracy of 77.257%. Lower CNN layers such as Conv1, 2 capture low-level features such as oriented edges and colors information. The higher layers such as Conv4 and Conv5 capture high-level semantic features. Based on these comparison results (Table 1), a deep feature extractor combining Conv4, Pooling2 and 5 layers was constructed.

4.2.3. DBNs architecture

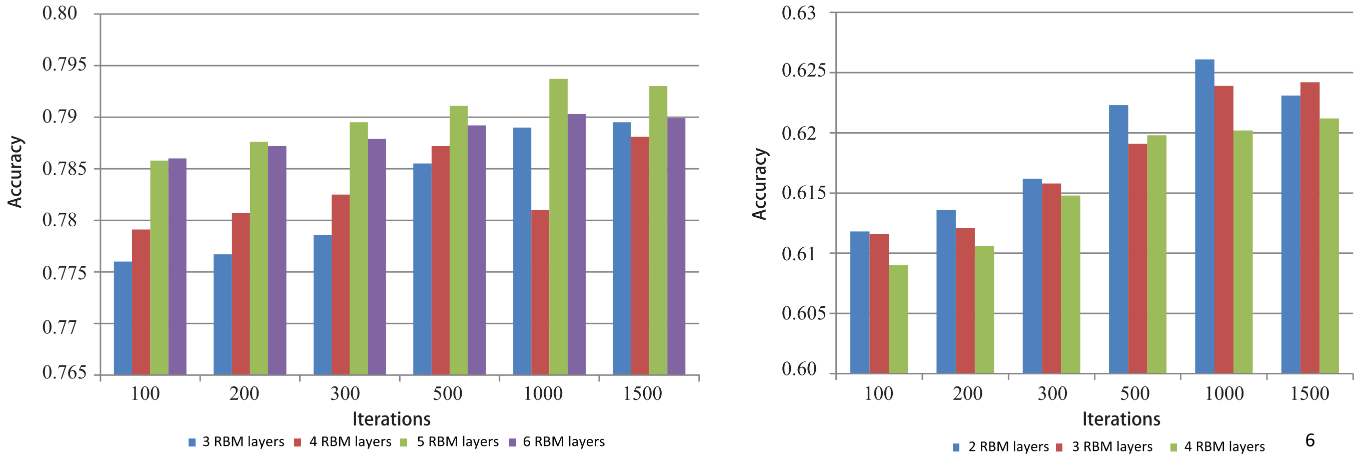
DBNs architecture. DBNs' perfromances are usually relevant to its architecture [53]. Several sets of experiments which tried different architectures of the DBNs for both shallow and deep modality channels were performed to study how the number of RBM layers in DBNs affects annotation accuracy. These experiments are conducted using the framework appearing in Fig. 2. For the deep modality channel, we fixed the DBNs for the shallow modality channel and tried four architectures: 3, 4, 5 and 6 for the RBM layer number, respectively. Fig. 10(a) shows the annotation results produced by these architectures. A small number of RBM layers generally leads to lower annotation accuracies. A large number of RBM layers (such as six layers) tend to over-fitting. DBNs was fixed to a deep modality channel and tried 2, 3, and 4 RBM layers to tune DBNs architecture in the shallow modality channel. Fig. 10(b) shows the corresponding results. Based on these experimental results, deep modality channels were set at 5 RBM layers and the DBNs in the shallow modality channel as 2 RBM layers. After all the above experiments, optimal parameter/architecture settings for all the modules in Fig. 2 were established. For both shallow and deep modality, as the iteration number increases, annotation precision increases gradually, and the training generally converges in 1000 epochs. Next the annotation performances of the different image representation schemes were compared.

4.2.4. Architecture comparison

To evaluate the multi-modal feature fusion scheme, five sets of experiments were designed. In the first experiment, the annotation with the experimental setting used in the architecture shown in Fig. 8(a) was performed but only considered the combination of SIFT, LBP, and Color features (i.e., the experimental setting which archives the best annotation accuracy in Section 4.2.1). The second to fourth experiments respectively used the MDMF, HSMF and HDMF features generated by the architecture in Fig. 8(b), (c) and (d). The last experiment used the fused multi-modality features generated by the architecture shown in Fig 2 (This architecture is named as MFF.).

A high-level, single-modality, features-based scheme achieves superior performance to mid-level, single-modality features-based ones (Table 2). DBNs plays an important role in improving image representation discrimination abilities. The proposed multi-modal feature fusion scheme achieves the best performance of all the five sets of experiments. It may be that the joint feature representation, which extracts both intrinsic and extrinsic properties of the objects in geographic images, discriminates strongly.

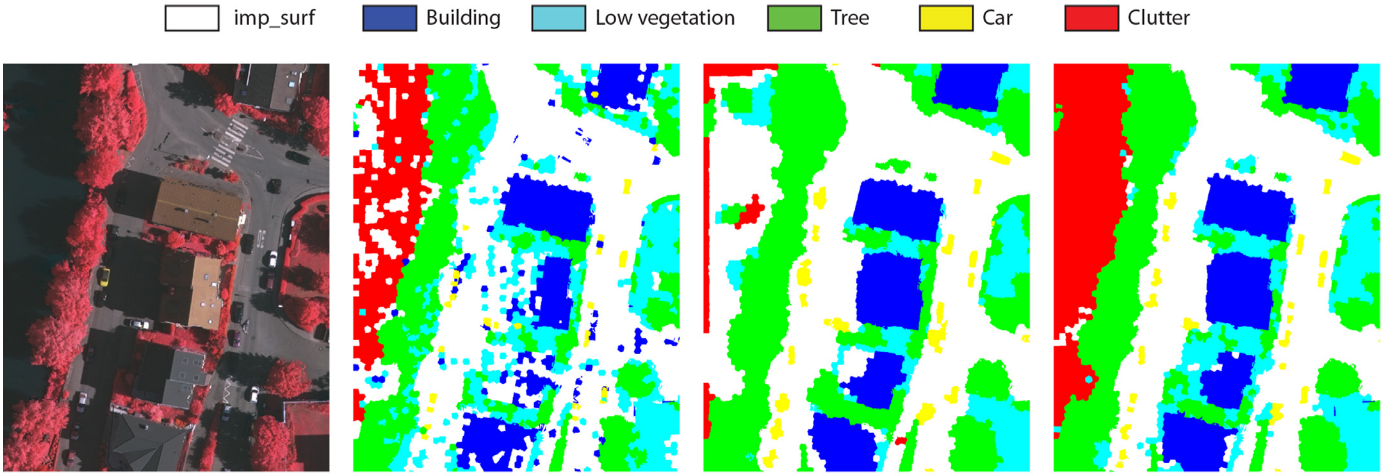
Tables 3 and 4 list the annotation accuracies of three schemes on a geographic image dataset. The results show that the proposed multi-modality fusion scheme achieved much better annotation performance than that using only single modality features in most of the six categories.



(a) deep modality

(b) shallow modality

Fig. 10. Comparison of annotation accuracies under different settings of the DBNs (Fig. 2(c)).



(a) original image

(b) shallow features

(c) deep features

(d) multi-modal features

Fig. 11. Annotation results produced by shallow feature combination, deep feature combination, and the proposed multi-modal on ISPRS benchmark.

4.3. Comparison with existing approaches

With the optimal framework determined, we now compare our proposed multi-modal feature fusion based approach with existing state-of-the-art techniques.

4.3.1. GeographicImages60CM300 Results

First, we compare the proposed multi-modal feature with two state-of-the-art methods: MvDN [36] and FCN [59] on the GeographicImages60CM300 dataset. For MvDN [59], mid-level shallow feature (LBP, SIFT, and Color features) and mid-level deep feature (Conv4; Pooling2, 5) as different views were used. For FCN [59], the AlexNet network as pre-train network and fine-tuning on the FCN network was used. The performances are measured by mean average precision (mAP) and standard deviation (std) on the test dataset.

Table 5 shows the corresponding results, FCN achieves a mAP of 79.2%. MvDN result is 80.3%. The proposed method achieves a mAP of 82.0%, 2.8 points higher than FCN and 1.7 points higher than MvDN. This may be due to the fusing of both deep and shallow features information, and mining correlations between high-level features from both shallow, and deep, modalities. Using the joint representation increases the intraclass similarity while reducing the interclass similarity.

Table 5
Comparison of annotation accuracies with different methods on GeographicImages60CM300 dataset.

	MvDN	FCN	MFF
Building	**78.9%** ± 3.29%	77.3% ± 4.03%	78.8% ± 1.52%
River	**88.8%** ± 3.01%	80.3% ± 3.37%	88.3% ± 2.17%
Road	60.4% ± 2.37%	60.2% ± 2.34%	**82.7%** ± 2.72%
Waste Land	80.0% ± 2.59%	82.4% ±3.83%	**82.8%** ± 1.91%
Fram Land	**96.1%** ± 2.13%	90.9% ±1.68%	79.1% ± 3.09%
Forest	77.5% ± 3.28%	**84.7%** ±2.33%	80.4% ± 3.26%
Mean Acc.(Std Acc)	80.3% ± 2.78%	79.3% ± 2.93%	**82.0%** ± 2.44%

4.3.2. ISPRS benchmark results

On ISPRS benchmark, the proposed method was compared to four other published state-of-the-art methods: Boost+CRF [38], CNN+CRF [44], FCN [59], and MvDN [36]. The annotation performances of Boost+CRF, CNN+CRF, and FCN used the results posted on the ISPRS website. MvDN had not yet been evaluated on ISPRS benchmark. Its performance was computed by ISPRS organizer based on our implementation. It is worth mentioning that most of annotation results posted on ISPRS website are produced by using two different kinds of dataset consisting of geospatial images and corresponding digital surface models (DSMs). All the comparison methods in this set of experiment only used geospatial images.

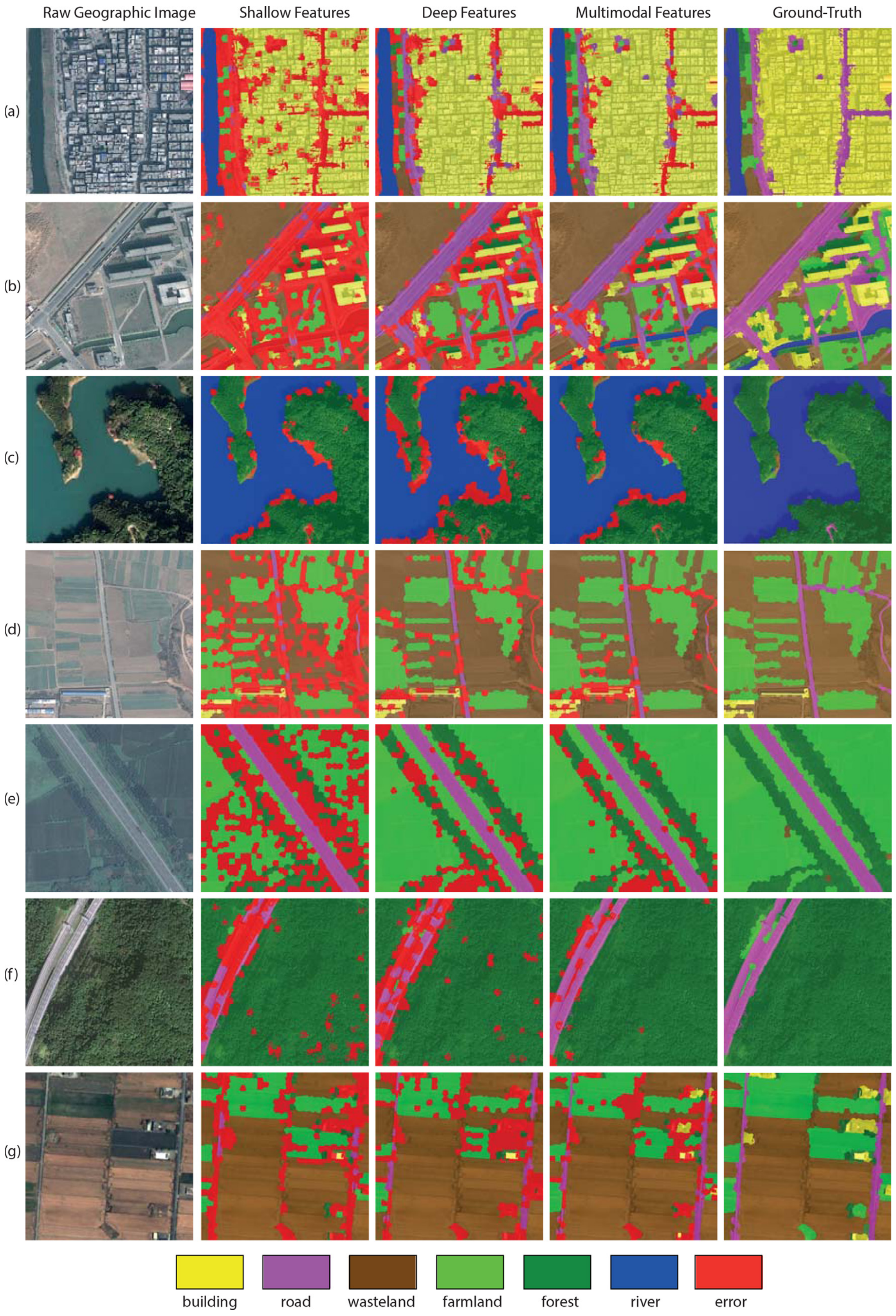


Fig. 12. Annotation results produced by HSMF, HDMF, and the proposed MFF on seven representative examples. Wrongly annotated pixels are marked in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 6
Performance comparisons of five different methods in term of mAP on ISPRS benchmark.

Method	Impsurf	Building	Lowveg	Tree	Car	Overall
MvDN [36]	87.5%	92.3%	71.0%	84.5%	67.3%	84.2%
Boost+CRF [38]	82.1%	82.8%	71.6%	81.6%	51.9%	79.4%
CNN+CRF [44]	82.9%	88.1%	67.1%	81.9%	53.5%	80.1%
FCN [59]	86.8%	90.8%	73.0%	84.6%	42.2%	84.1%
MFF	**90.7%**	**94.4%**	**81.3%**	**88.2%**	**74.8%**	**88.7%**

Table 6 shows the quantitative comparison results of five different methods. The proposed method outperforms all other approaches in terms of AP. Shallow and deep modality feature fusion boosts annotation performances. Fig. 11 shows ISPRS benchmark annotation results. Column 1 contains the original geographic images. Columns 2–4 show annotation results using the single-modality shallow features, single-modality deep features, and the fused features. Clutter is better annotated by the shallow features than by deep features, which shows the advantage of shallow modality features on simple geospatial objects and shallow modality features can provide complementary information for deep modality features, (Fig. 11 (b) and (c)).

4.3.3. *Qualitative analysis*

In Fig. 12, the annotation results of 7 representative images in our experiments are shown. Column 1 contains the original geographic images. Columns 2–4 show annotation results using respectively HSMF, HDMF, and MFF features. and Column 5 is the ground-truth. The visual results indicate that, for most superpixels of examples, the annotations produced by using deep features are more accurate than those generated by shallow features, and the fused features achieves the best results.

Specifically, river annotation results produced by HSMF have very low accuracies. See Row 2 and Column 2 of Fig. 12. All the superpixels of the river were recognized as farm land because the color of the river is very close to that of the farm land. However, HSMF exceeds HDMF on the forest in Fig. 12(f) and the lake (river) in Fig. 12(c), showing the advantages of shallow-modality features on simple geospatial objects. This may result from the fact that shallow features represent extrinsic visual properties of geographic images.

5. Conclusions and future work

A novel multi-modal feature fusion method is presented in this paper for learning a discriminative image representation for geographic image annotation tasks. The method contains two feature extraction channels: one for shallow feature extraction, whereas the other for deep feature extraction. Each channel has a low-to-high level feature learning flow. It first extracts low-level features, including shallow-modality features (SIFT, Color, and LBP) and deep-modality features (CNNs) for each pixel of the input image, then constructs mid-level features from the low-level features of the individual pixels into superpixels. Finally it learns high-level features from the mid-level features by using deep belief networks (DBNs). A restricted boltzmann machine (RBM) is used to mine correlations between high-level features from both shallow and deep modalities. It produces the final representation for input geographic images to fuse high level shallow and deep features.

By comprehensive experiments on various image representation methods, the following conclusions are proposed:

- single-modality deep feature based method can archive better annotation accuracy than methods based on a single-modality shallow feature combination for most geospatial objects;
- methods based on multi-modal feature fusion can achieve better performance than single-modality feature based method;
- shallow features can be complementary to deep features.

The aim of the proposed method is to fuse multi-modal features for improving discriminative capability of extracted features. Spatial relationships are not considered, and structural relations of objects in the image are not utilized. According to related research works [60,61], through structural learning, the image parsing accuracy can be improved. In the following research, adapting long short term memory (LSTM) was considered to learn structural relationship of superpixel. That is, extracted features will be fused into high-dimensional features, from which the features for superpixels are computed through structural learning using a graphical LSTM.

Acknowledgments

We thank all the reviewers for their valuable comments and feedback. This work is supported in part by grants from the China's National Key R&D Program (2017YFB0503503), National Natural Science Foundation of China (61573284), the Science and Technology Plan Project of Hunan Province (2016TP1020), HNU Start-up Grant (15B22), and the Program of Key Disciplines in Hunan Province.

References

[1] D. Lu, Q. Weng, A survey of image classification methods and techniques for improving classification performance, Int. J. Remote Sens. 28 (5) (2007) 823–870.
[2] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[3] J.G. Daugman, Two-dimensional spectral analysis of cortical receptive field profiles, Vis. Res. 20 (10) (1980) 847–856.
[4] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition, 1, 2005, pp. 886–893.
[5] Y. Bengio, A. Courville, P. Vincent, Representation learning: a review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1798–1828.
[6] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, T. Darrell, Decaf: a deep convolutional activation feature for generic visual recognition, arXiv preprint arXiv:1310.1531 (2013).
[7] A. Razavian, H. Azizpour, J. Sullivan, S. Carlsson, CNN features off-the-shelf: an astounding baseline for recognition, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops, 2014, pp. 806–813.
[8] H. Aytekin Zongur, Texture-based airport runway detection, Geosci. Remote Sens. Lett. 10 (3) (2013) 471–475.
[9] Y. Yang, S. Newsam, Geographic image retrieval using local invariant features, IEEE Trans. Geosci. Remote Sens. 51 (2) (2013) 818–832.
[10] L. Chen, W. Yang, K. Xu, T. Xu, Evaluation of local features for scene classification using vhr satellite images, in: Proc. Urban Remote Sensing Event, 2011, pp. 385–388.
[11] T. Ahonen, A. Hadid, M. Pietikainen, Face description with local binary patterns: application to face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 28 (12) (2006) 2037–2041.
[12] C. Tao, Y. Tan, H. Cai, J. Tian, Airport detection from large ikonos images using clustered sift keypoints and region information, Geosci. Remote Sens. Lett. 8 (1) (2011) 128–132.
[13] Y. Li, X. Sun, H. Wang, H. Sun, X. Li, Automatic target detection in high-resolution remote sensing images using a contour-based spatial model, Geosci. Remote Sens. Lett. 9 (5) (2012) 886–890.
[14] J. Li, H. Zhang, L. Zhang, X. Huang, L. Zhang, Joint collaborative representation with multitask learning for hyperspectral image classification, IEEE Trans. Geosci. Remote Sens. 52 (9) (2014) 5923–5936.
[15] Y. Yang, S. Newsam, Semi-supervised learning of geospatial objects through multi-modal data integration, in: Proc. 22nd International Conference on Pattern Recognition, 2014, pp. 4062–4067.
[16] L. Zhang, L. Zhang, D. Tao, X. Huang, On combining multiple features for hyperspectral remote sensing image classification, IEEE Trans. Geosci. Remote Sens. 50 (3) (2012) 879–893.
[17] D. Tuia, G. Camps-Valls, Urban image classification with semisupervised multiscale cluster kernels, Sel. Top. Appl. Earth Obs. Remote Sens. 4 (1) (2011) 65–74.
[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2009, pp. 248–255.

[19] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Proc. Advances in neural information processing systems, 2012, pp. 1097–1105.

[20] L.C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Semantic image segmentation with deep convolutional nets and fully connected CRFs, in: Proc. International Conference on Learning Representations, 2015.

[21] W. Luo, H. Li, G. Liu, Automatic annotation of multispectral satellite images using author–topic model, Geosci. Remote Sens. Lett. 9 (4) (2012) 634–638.

[22] B. Sirmacek, C. Ünsalan, Urban-area and building detection using sift keypoints and graph theory, IEEE Trans. Geosci. Remote Sens. 47 (4) (2009) 1156–1167.

[23] H. Grabner, T.T. Nguyen, B. Gruber, H. Bischof, On-line boosting-based car detection from aerial images, ISPRS J. Photogramm. Remote Sens. 63 (3) (2008) 382–396.

[24] L. Wang, S. Hao, Q. Wang, Y. Wang, Semi-supervised classification for hyperspectral imagery based on spatial-spectral label propagation, ISPRS J. Photogramm. Remote Sens. 97 (2014) 123–137.

[25] C. Farabet, C. Couprie, L. Najman, Y. LeCun, Learning hierarchical features for scene labeling, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1915–1929.

[26] J. Yue, W. Zhao, S. Mao, H. Liu, Spectral spatial classification of hyperspectral images using deep convolutional neural networks, Remote Sens. Lett. 6 (6) (2015) 468–477.

[27] W. Zhao, Z. Guo, J. Yue, X. Zhang, L. Luo, On combining multiscale deep learning features for the classification of hyperspectral remote sensing imagery, Int. J. Remote Sens. 36 (13) (2015) 3368–3379.

[28] X. Chen, S. Xiang, C.-L. Liu, C.-H. Pan, Vehicle detection in satellite images by hybrid deep convolutional neural networks, Geosci. Remote Sens. Lett. 11 (10) (2014) 1797–1801.

[29] X. Huang, L. Zhang, Comparison of vector stacking, multi-svms fuzzy output, and multi-svm+s voting methods for multiscale vhr urban mapping, Geosci. Remote Sens. Lett. 7 (2) (2010) 261–265.

[30] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A.Y. Ng, Multimodal deep learning, in: Proc. the 28th International Conference on Machine Learning, 2011, pp. 689–696.

[31] N. Srivastava, R.R. Salakhutdinov, Multimodal learning with deep boltzmann machines, in: Proc. Advances in neural information processing systems, 2012, pp. 2222–2230.

[32] J. Li, H. Zhang, L. Zhang, Efficient superpixel-level multitask joint sparse representation for hyperspectral image classification, IEEE Trans. Geosci. Remote Sens. 53 (10) (2015) 5338–5351.

[33] X. Zheng, X. Sun, K. Fu, H. Wang, Automatic annotation of satellite images via multifeature joint sparse coding with spatial relation constraint, Geosci. Remote Sens. Lett. 10 (4) (2013) 652–656.

[34] A. Shahroudy, T.-T. Ng, Y. Gong, G. Wang, Deep multimodal feature analysis for action recognition in rgb+d videos, arXiv preprint arXiv:1603.07120 (2016).

[35] M. Kan, S. Shan, X. Chen, Multi-view deep network for cross-view classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4847–4855.

[36] W. Wang, R. Arora, K. Livescu, J.A. Bilmes, On deep multi-view representation learning., in: ICML, 2015, pp. 1083–1092.

[37] J. Porway, K. Wang, B. Yao, S.C. Zhu, A hierarchical and contextual model for aerial image understanding, in: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE, 2008, pp. 1–8.

[38] I. Markus Gerke, Use of the stair vision library within the isprs 2d semantic labeling benchmark (vaihingen).

[39] A.M. Cheriyadat, Unsupervised feature learning for aerial scene classification, IEEE Trans. Geosci. Remote Sens. 52 (1) (2014) 439–451.

[40] A. Kembhavi, D. Harwood, L.S. Davis, Vehicle detection using partial least squares, IEEE Trans. Pattern Anal. Mach. Intell. 33 (6) (2011) 1250–1265.

[41] Y. Chen, Z. Lin, X. Zhao, G. Wang, Y. Gu, Deep learning-based classification of hyperspectral data, Sel. Top. Appl. Earth Obs. Remote Sens. 7 (6) (2014) 2094–2107.

[42] M. Castelluccio, G. Poggi, C. Sansone, L. Verdoliva, Land use classification in remote sensing images by convolutional neural networks, arXiv preprint arXiv:1508.00092 (2015).

[43] D. Marmanis, M. Datcu, T. Esch, U. Stilla, Deep learning earth observation classification using imagenet pretrained networks, IEEE Geosci. Remote Sens. Lett. 13 (1) (2016) 105–109.

[44] S. Paisitkriangkrai, J. Sherrah, P. Janney, V.-D. Hengel, et al., Effective semantic pixel labelling with convolutional networks and conditional random fields, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2015, pp. 36–43.

[45] J. Sherrah, Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery, arXiv preprint arXiv:1606.02585 (2016).

[46] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Susstrunk, Slic superpixels compared to state-of-the-art superpixel methods, IEEE Trans. Pattern Anal. Mach. Intell. 34 (11) (2012) 2274–2282.

[47] A. Vedaldi, B. Fulkerson, Vlfeat: An open and portable library of computer vision algorithms, in: Proc. 18th ACM International Conference on Multimedia, 2010, pp. 1469–1472.

[48] T. Ojala, M. Pietikainen, D. Harwood, A comparative study of texture measures with classification based on featured distributions, Pattern Recognit. 29 (1) (1996) 51–59.

[49] X. Wang, T.X. Han, S. Yan, An HOG-LBP human detector with partial occlusion handling, in: Proc. 12th International Conference on Computer Vision, 2009, pp. 32–39.

[50] I.T. Jolliffe, Principal component analysis, 2005.

[51] H. Larochelle, D. Erhan, A. Courville, J. Bergstra, Y. Bengio, An empirical evaluation of deep architectures on problems with many factors of variation, in: Proc. 24th International Conference on Machine Learning, 2007, pp. 473–480.

[52] H. Lee, R. Grosse, R. Ranganath, A.Y. Ng, Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations, in: Proc. 26th International Conference on Machine Learning, 2009, pp. 609–616.

[53] G. Hinton, A practical guide to training restricted boltzmann machines, in: Neural Networks: Tricks of the Trade, 2012, pp. 599–619.

[54] Y. Bengio, Learning deep architectures for AI, Found. Trends Mach. Learn. 2 (1) (2009) 1–127.

[55] G.E. Hinton, S. Osindero, Y.-W. Teh, A fast learning algorithm for deep belief nets, Neural Comput. 18 (7) (2006) 1527–1554.

[56] K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 37 (9) (2015) 1904–1916.

[57] F. Wilcoxon, Individual comparisons by ranking methods, Biom. Bull. 1 (6) (1945) 80–83.

[58] (http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html).

[59] (http://ftp.ipi.uni-hannover.de/ISPRS_WGIII_website/ISPRSIII_4_Test_results/papers/Description_UCalgary.pdf).

[60] W. Byeon, T.M. Breuel, F. Raue, M. Liwicki, Scene labeling with lstm recurrent neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3547–3555.

[61] X. Liang, X. Shen, J. Feng, L. Lin, S. Yan, Semantic object parsing with graph lstm, in: European Conference on Computer Vision, Springer, 2016, pp. 125–143.

Ke Li received the PhD degree in Zhengzhou Institute of Surveying and Mapping, China in 2008. He is an associate professor at Zhengzhou Institute of Surveying and Mapping, China. His research includes GIS, deep learning, image processing, and computer vision. He has published over 15 journal and conference papers in the related areas.

Changqing Zou received the B.E. degree from Harbin Institute of Technology, China, in 2005, and the M.E. degree from Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, China, in 2008, and the PHD degree at the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China, in 2015. He is now a post doctoral research fellow in Simon Fraser University, Canada. His research interests include computer vision and computer graphics.

Shuhui Bu received the MSc and PhD degrees in the College of Systems and Information Engineering from University of Tsukuba, Japan in 2006 and 2009. He was an assistant professor (2009–2011) at Kyoto University, Japan. Currently, he is an associate professor at Northwestern Polytechnical University, China. His research interests are concentrated on computer vision and robotics, including SLAM, 3D shape analysis, image processing, pattern recognition, 3D reconstruction, and related fields. He has published approximately 40 papers in major international journals and conferences, including the IEEE TMM, IEEE TBME, TVC, C&G, ACM MM, ICPR, CGI, SMI, etc.

Yun Liang is an associate professor and Master advisor of College of Information in South China Agricultural University and a part-time associate research fellow in the “National Engineering Research Center of Digital Life” of China. She was born in 1981 in Shandong Province. She received her M.S. and PHD degree from the “National Engineering Research Center of Digital Life” in Information Science and Technology of Sun Yat-sen University in 2005, 2011 respectively. Her research interests include image processing, compute vision and machine learning. She has published 22 academic papers and applied for two national patients.

Jian Zhang received his B.E. degree and M.E. degree from Shandong University of Science and Technology in 2000 and 2003 respectively, and received the Ph.D. degree from Zhejiang University in 2008. He is an associate professor working in School of Science & Technology of Zhejiang International Studies University. Before this, he worked in the department of mathematics of Zhejiang University from 2009 to 2011. His research interests include computer animation, multimedia processing and machine learning. He serves as a reviewer of several prestigious journals.

Minglun Gong is a professor at the Memorial University of Newfoundland and an adjunct professor at the University of Alberta. He obtained his Ph.D. from the University of Alberta in 2003, his M.Sc. from the Tsinghua University in 1997, and his B.Engr. from the Harbin Engineering University in 1994. After graduation, he was a faculty member at the Laurentian University for four years before joined the Memorial University. Minglun’s research interests cover various topics in the broad area of visual computing (including computer graphics, computer vision, visualization, image processing, and pattern recognition). So far, he has published over 100 referred technical papers in journals and conference proceedings, including 15 papers in ACM/IEEE transactions. He is the inventor of an awarded patent and 3 pending patents. Currently an associate editor for Pattern Recognition, he has also served as program committee member for top-tier conferences (e.g. ICCV and CVPR) and reviewer for prestigious journals (e.g. IEEE TPAMI and ACM TOG). He was the recipient of the Izaak Walton Killam Memorial Award and the CFI New Opportunity Award.