Journal of
Applied Remote Sensing

RemoteSensing.SPIEDigitalLibrary.org

Deep learning decision fusion for the classification of urban remote sensing data

Ghasem Abdi
Farhad Samadzadegan
Peter Reinartz

SPIE.

Ghasem Abdi, Farhad Samadzadegan, Peter Reinartz, "Deep learning decision fusion for the classification of urban remote sensing data," *J. Appl. Remote Sens.* **12**(1), 016038 (2018), doi: 10.1117/1.JRS.12.016038.

Deep learning decision fusion for the classification of urban remote sensing data

Ghasem Abdi,[a,*] Farhad Samadzadegan,[a] and Peter Reinartz[b]
[a]University of Tehran, College of Engineering, Faculty of Surveying and Geospatial Engineering, Tehran, Iran
[b]German Aerospace Centre (DLR), Remote Sensing Technology Institute, Department of Photogrammetry and Image Analysis, Weßling, Germany

Abstract. Multisensor data fusion is one of the most common and popular remote sensing data classification topics by considering a robust and complete description about the objects of interest. Furthermore, deep feature extraction has recently attracted significant interest and has become a hot research topic in the geoscience and remote sensing research community. A deep learning decision fusion approach is presented to perform multisensor urban remote sensing data classification. After deep features are extracted by utilizing joint spectral–spatial information, a soft-decision made classifier is applied to train high-level feature representations and to fine-tune the deep learning framework. Next, a decision-level fusion classifies objects of interest by the joint use of sensors. Finally, a context-aware object-based postprocessing is used to enhance the classification results. A series of comparative experiments are conducted on the widely used dataset of 2014 IEEE GRSS data fusion contest. The obtained results illustrate the considerable advantages of the proposed deep learning decision fusion over the traditional classifiers. © *2018 Society of Photo-Optical Instrumentation Engineers (SPIE)* [DOI: 10.1117/1.JRS.12.016038]

Keywords: convolutional neural network; decision-level fusion; deep features; deep learning; stacked sparse autoencoder; thermal hyperspectral.

Paper 171018 received Nov. 28, 2017; accepted for publication Feb. 12, 2018; published online Mar. 13, 2018.

1 Introduction

With the recent tremendous advances in remote sensing imaging systems, there has been a significant increase in making use of well-defined multiple sensors and sources via the geoscience and remote sensing research community to achieve a robust and complete description about objects of interest.[1–5] In this respect, image analysis and data fusion play a significant role in applications of pansharpening,[6,7] classification,[8–10] change detection,[11,12] large-scale processing,[13,14] multiple resolutions,[15,16] domain adaption,[17–19] interactive systems,[20,21] and signal-level fusion with different meanings and properties.[22] In the above-mentioned applications, classification of remote sensing images from urban area is one of the most challenging topics, which are still ongoing and have potential in further development due to a wide variety of man-made objects and scene complexity so that urban area classification has attracted considerable interest and has turned into a hot research topic in the geoscience and remote sensing research community.[5,23] Within this context, fusion of aerial visible and thermal hyperspectral data has been received increasing interest over the past few years.[5,23–28] Aerial visual data (with considerable spatial descriptors) play a significant role in the urban land cover classification and demonstrate important spectral features in the visible spectrum.[24] Furthermore, the advances in thermal imaging technology have made it possible to be collected in multiple continuous spectral channels with significantly improved joint spectral–spatial resolutions for identification of various physical materials regardless of illumination conditions, leading thereby to enhanced

*Address all correspondence to: Ghasem Abdi, E-mail: ghasem.abdi@ut.ac.ir

1931-3195/2018/$25.00 © 2018 SPIE

classification performance.[24,29,30] The development of spectral-based classification of thermal remote sensing data has focused on spectral absorption descriptors of silicate minerals, the main parts of the terrestrial surface, and man-made construction objects. The silicon–oxygen bonds of the silicate minerals cannot present distinct spectral descriptors in the visible-to-short-wave infrared range,[31] whereas its stretching vibrations display considerable spectral features in the longwave infrared range.[32] In this case, man-made objects emit a greater extent of polarized infrared radiation than naturally derived background materials[33,34] because they have relatively smooth surface features versus most naturally occurring surfaces. Within this context, the emissivity can parametrically suffice if surface irregularities are large, relative to the emitted radiation's wave range, whereas the surface may be more specular and an observable induced polarization arises in the emitted thermal radiation if surface irregularities are small versus the emission wave range.[35]

Classification of urban visual and thermal hyperspectral data is a new subject in the geoscience and remote sensing research community and limited research have been conducted in this field of study. Liao et al.[24] presented the outcomes of 2014 IEEE GRSS data fusion contest, a challenging multiresolution and multisensor image analysis and data fusion research community problem. The winning article of the classification section focused on hierarchical classification strategy to combine visual and thermal hyperspectral data. In this context, the land cover classes are successively identified by a binary support vector machine (SVM) classifier on the concatenated feature descriptors. In addition, the obtained pixel-based land cover classification map is refined by majority voting, adaptive mean shift segmentation and multiple semantic rules. The winners of the paper contest provided a development of the multiresolution and multisensor image analysis and data fusion. In this respect, visual data are applied in a guided filtering scheme to improve thermal image's spatial resolution, and then the land cover classification map is identified using an SVM classifier on a supervised graph-based feature-level fusion. After the contest, the data are publicly available still as a challenging image analysis opportunity for further development. Lu et al.[5] presented a decision-level classifier fusion to identify a thematic land cover classification map of the 2014 IEEE GRSS data fusion contest dataset. A semisupervised local discriminant analysis extracts distinct thermal feature descriptors that are fed into an SVM classifier, followed by feature representations of joint spectral–textural information to perform land cover classification of aerial visible data. In addition, an object-based decision-level fusion is proposed to integrate the above-mentioned classifiers and to enhance the classification performance. Li et al.[23] suggested a multilevel land cover classification approach to integrate the 2014 IEEE GRSS data fusion contest dataset. In this case, road pixels are first identified by combination of thermal imagery classification outcomes and visual data segmentation map, and then the rest of classes are classified by utilizing joint spectral–spatial information of visual data. Moreover, an object-based decision-level fusion is applied to enhance the classification performance. Eslami and Mohammadzadeh[25] applied in-scene atmospheric compensation to thermal hyperspectral data, and then sequential parametric projection pursuit dimensionality reduction is employed to overcome finite training set problem with the high dimensionality of hyperspectral data. Furthermore, an SVM classifier (on a visual and the above-mentioned thermal feature-level fusion) is employed to identify the land cover classes, and eventually an object rule-based postprocessing refines the obtained pixel-based land cover classification map. Samadzadegan et al.[27] proposed a cuckoo search optimization algorithm with mixed binary-continuous coding to determine a suitable subset of feature representations of joint spectral–spatial information and SVM hyperparameters simultaneously. Eslami and Mohammadzadeh[26] presented a hierarchical classification strategy to integrate visual spectral–spatial and thermal spectral–textural feature descriptors. In this context, the urban land cover mapping is successively identified using a binary SVM classifier on the above-mentioned feature descriptors. In addition, the obtained pixel-based land cover classification map is improved by an object rule-based postprocessing. Abdi et al.[28] presented a decision-level classifier fusion strategy to combine the 2014 IEEE GRSS data fusion contest dataset. In this context, road pixels are first classified by thermal spectral information, and then the rest of classes are successively identified by the joint use of sensors via Dempster–Shafer classifier fusion. Furthermore, an object-based postprocessing (OBP) is used to enhance the classification performance. In the above-described papers, many ways were presented to classify the 2014

IEEE GRSS data fusion contest dataset; they provided fascinating novelty in development and practical classification applications. Within this context, two common dimensionality reduction strategies were widely applied to overcome finite training set problems with the high dimensionality of hyperspectral data. The dimensionality reduction by transform employs a transformation function to obtain some optimum through data compression, whereas the band selection of dimensionality reduction exploits an appropriate subset of spectral bands to reduce input space by a definite optimization criterion.[36] Moreover, spectral–spatial remote sensing data classification were extensively investigated to enhance the classification performance by considering homogeneous areas as a set of neighboring pixels whose spectral information are mainly belonging to one class.[37–39] Most existing techniques defined shallow handcrafted features or transform-based filters of the input data that are not robust enough to make a deal with classification challenges of remote sensing data.[40] Additionally, deep learning frameworks have recently enhanced the classification performance by automatic extraction of extremely powerful deep features and therefore has led to a hot research topic in the geoscience and remote sensing research community;[41–53] i.e., the traditional shallow architectures are replaced by novel deep frameworks motivated by the human brain architectural model.[54] From the deep learning point of view, deep belief networks train one layer using restricted Boltzmann machines in an unsupervised manner.[55,56] Autoencoder and its variants learn the intermediate layers of representation in an unsupervised way.[57,58] Unlike autoencoders, the sparse coding methods exploit sparse representations of the input space via training a dictionary.[59] Furthermore, convolutional neural networks, the most representative supervised deep learning framework, admit the deep architecture to train invariant and abstract features and to convert the original data into representations that can notably enhance the classification performance.[60] The deep learning methods are explained in great details in the machine learning research literature.[61,62]

Deep learning turns out to be a significant developing trend in the geoscience and remote sensing research community and remains extremely challenging due to its novelty and limited research up to now. In this paper, we propose a deep learning decision fusion for the classification of urban remote sensing data and address the advantages of the presented method over a large number of traditional classifiers.

2 Proposed Method

In this article, we present a deep learning decision fusion for the classification of urban remote sensing data that contain deep feature extraction, logistic regression classifier, decision-level classifier fusion, and context-aware OBP steps. A deep architecture is designed to progressively learn invariant and abstract feature representations of the input data comprised of spectral, spatial, or joint utilization of spectral–spatial features, and then a logistic regression classifier is applied to train high-level features at the top layer and to optimize the deep learning framework. In the next step, an enhanced classification map is estimated by integrating multiple classifier outcomes, followed by a context-aware OBP refining the obtained pixel-based land cover classification map. The general structure of the presented method is shown in Fig. 1.

2.1 *Deep Feature Extraction*

The joint spectral–spatial classification framework is constructed by the concatenated spectral–spatial feature descriptors. In this case, the raw spectral data are first used to take advantage of the available contiguous spectral bands for classification applications. Second, the spatial feature descriptors are made by a local window considering that a set of neighboring pixels for many cases belong to one class. Finally, a hybrid set of joint spectral–spatial information is made by stacking the above-mentioned feature descriptors.

2.1.1 *Stacked sparse autoencoder*

A shallow sparse autoencoder exhibits a specific type of neural network consisting input, hidden and reconstruction layers that are used to progressively learn invariant and abstract features in an

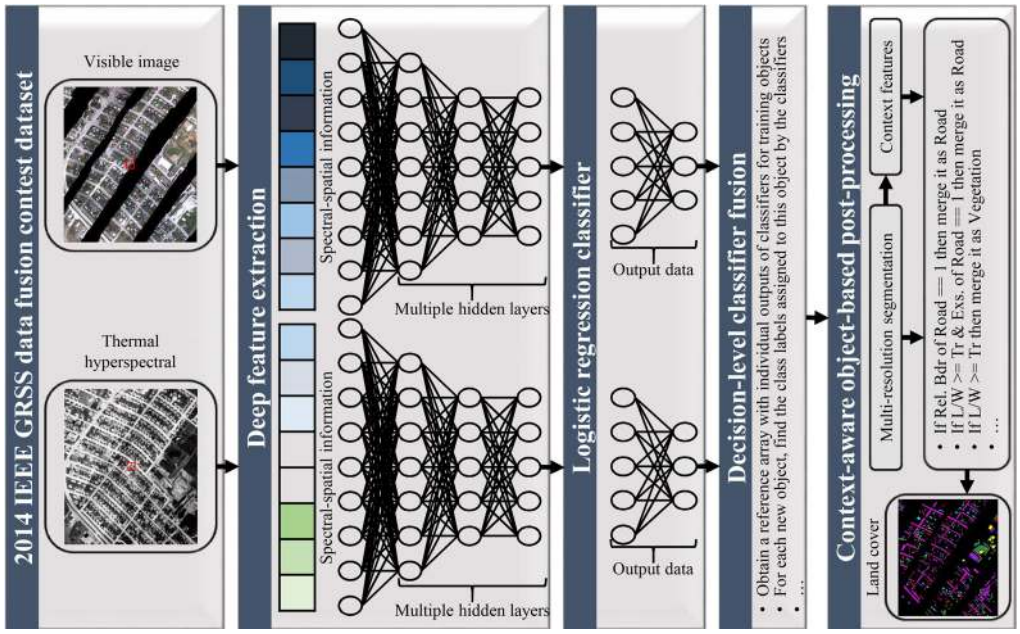

Fig. 1 Flowchart of the proposed method.

unsupervised manner;[63,64] i.e., an encoder function transfers the input-to-hidden layer, and then a decoder function reconstructs an approximation of the input data from hidden representation by

$$z = f(W_z x + b_z), \tag{1}$$

$$y = f(W_y z + b_y), \tag{2}$$

where W_z and W_y define the input-to-hidden and the hidden-to-output weights, whereas b_z and b_y indicate the bias of the hidden and output units. In addition, f is the logistic sigmoid function, defining nonlinear mapping function of the encoder and decoder transitions. The optimal parameters are commonly estimated by minimizing the reconstruction error with sparsity constraint and weight decay terms as

$$J_{\text{cost}} = \frac{1}{M}\sum_{i=1}^{M}\left(\frac{1}{2}\|y_i - x_i\|_2^2\right) + \frac{\lambda}{2}\sum_{l}\sum_{i}\sum_{j}[W_{i,j}^{(l)}]^2 + \eta\sum_{j=1}^{S} KL(r\|\overline{r}_j), \tag{3}$$

in the above equation, the initial term defines the reconstruction error of M training samples,[43] the second one explains the weight decay term to minimize the over fitting of autoencoder,[65] and the last one is a sparsity penalty term to enforce the average latent unit activation to be close to the target value.[43] Moreover, the minimization strategy of the above-described function can be iteratively carried out utilizing stochastic gradient descent and back propagation method.[66,67]

A deep stacked sparse autoencoder (SAE) is developed to progressively train high-level feature representations of input data by stacking several layers of sparse autoencoders and can be trained utilizing a greedy layerwise strategy for extra layers.

2.1.2 *Convolutional neural network*

Convolutional networks present a specific type of neural network containing an input, multiple hidden, and an output layer that is used to progressively learn invariant and abstract features in a supervised manner. A typical convolutional network employs a convolutional layer with subsequent nonlinear operations and a pooling layer.[40] A convolutional layer can be defined as

$$x_j^l = f\left(\sum_{i=1}^{M} x_i^{l-1} * k_{ij}^l + b_j^l\right), \tag{4}$$

where x_i^{l-1} defines the i'th feature map of $l-1$'th layer, x_j^l explains the j'th feature map of l'th layer, and M indicates the number of input feature maps. k_{ij}^l and b_j^l exhibit the trainable parameters in the convolutional layer. f represents a function that enhances the nonlinear properties of the decision function and of the entire network without influencing the convolution layer's receptive fields.

The pooling layer is applied to reduce the spatial resolution of the representation, the number of parameters, and computation time in the network by dividing the input data into a set of nonoverlapping rectangles and results in the maximum of each subregion.

A deep convolutional neural network (CNN) is designed to progressively train high-level feature representations of input data by stacking multiple convolutional layers followed by nonlinear mapping function and different pooling layers.

2.2 *Logistic Regression Classifier*

After deep feature extraction, a logistic regression classifier is augmented above the output feature descriptors of the highest network's layer to fine-tune the deep learning architecture by enforcing gradient descent from the current setting of the parameters to minimize the training error on the labeling samples. Within this context, softmax regression represents an extended type of logistic regression that can be utilized for multiclass classification targets, and its outcome can be defined as a set of conditional probabilities by

$$P(Y = i|R, W, b) = s(WR + b) = \frac{e^{W_i R + b_i}}{\sum_j e^{W_j R + b_j}}, \tag{5}$$

where R indicates a result of the network's last hidden layer, W and b denote the logistic regression layer's weights and biases. The above-described fine-tuning is generally performed by defining very slight learning rates on the network's layers.[42]

2.3 *Decision-Level Classifier Fusion*

Classifier fusion has been used on several types of data to enhance the performance of individual classifiers. Within this context, a set of decisions is first conducted, and next synthesized via a particular classifier fusion strategy; the combined decision commonly demonstrates more precise, accurate, and certain toward any decisions that construct the ensemble.[68–70]

2.3.1 *Naïve Bayes combination*

Naïve Bayes (NB) combination, one of the most efficient classifier fusion strategies, can be applied to integrate the classifier label outputs. Denote $P(s_j)$ the probability that classifier D_j labels x in class $s_j \in \Omega$ the conditional independence can be defined by

$$P(s|\omega_k) = P(s_1, s_2, \ldots, s_L|\omega_k) = \prod_{i=1}^{L} P(s_i|\omega_k), \tag{6}$$

where L defines the number of classifiers. The posterior probability required to classify x can be explained by

$$P(\omega_k|s) = \frac{P(\omega_k)P(s|\omega_k)}{P(s)} = \frac{P(\omega_k)\prod_{i=1}^{L} P(s_i|\omega_k)}{P(s)}, \qquad k = 1, \ldots, c, \tag{7}$$

where c indicates the number of classes. The denominator does not rely on ω_k and can be disregarded, so the support for class ω_k is determined by

$$\mu_k(x) \propto P(\omega_k) \prod_{i=1}^{L} P(s_i|\omega_k). \tag{8}$$

The maximum result of μ indicates winner label for x instance. More detailed descriptions about the NB combination method can be found in Ref. 69.

2.3.2 *Behavior knowledge space combination*

The behavior knowledge space (BKS) combination strategy can efficiently integrate the classifier labels and extract more accurate outputs for the classification applications. The BKS combination method contains knowledge-modeling and operation steps. In the knowledge-modeling step, it exploits knowledge from *a priori* behavior of classifiers and constructs a K-dimensional BKS, and the operation stage is then conducted for all testing samples and combines individual classifier label outputs into a final decision by a rule that utilizes the knowledge inside of the unit. More detailed descriptions about the BKS combination method can be found in Ref. 68.

2.4 *Context-Aware Object-Based Postprocessing*

The above-mentioned pixel-based land cover classification map always has a large number of outlier classified pixels caused by the problem of excessive heterogeneity.[28] In this case, a multiresolution image segmentation method can be employed to split the image data into multiple

Table 1 Context-aware object-based postprocessing rules.

Majority vote	Relationships among spatial objects
Road	If relative border of road class $\neq 0$, then merge it into road object.
	If relative border of road class $= 0$, then merge it into adjacent image object.
Tree	If length to width $\leq$ TLW or area pixels $\leq$ TPX, then merge it into tree object.
	If length to width > TLW or area pixels > TPX, then merge it into vegetation object.
Red roof	If length to width $\leq$ TLW, then merge it into red roof object.
	If length to width > TLW and relative border of road class $\neq 0$, then merge it into road object.
	If length to width > TLW and relative border of road class = 0, then merge it into adjacent image object.
Gray roof	If length to width $\leq$ TLW, then merge it into gray roof object.
	If length to width > TLW and relative border of road class $\neq 0$, then merge it into road object.
	If length to width > TLW and relative border of road class = 0, then merge it into adjacent image object.
Concrete roof	If length to width $\leq$ TLW, then merge it into concrete roof object.
	If length to width > TLW and relative border of road class $\neq 0$, then merge it into road object.
	If length to width > TLW and relative border of road class = 0, then merge it into adjacent image object.
Vegetation	Merge it into vegetation object.
Bare soil	Merge it into bare soil object.
Land cover	If relative border of road class = 1, then merge it into road object.

spatially nonoverlapping regions. In this context, the multiresolution image segmentation method takes one pixel and constantly creates larger ones by considering a local homogeneity criterion between adjacent image objects, and also merging that pair of image objects.[71] After completing multiresolution image segmentation, the final label of each segmented region is made by majority voting and considering relationships among spatial objects (Table 1).

3 Experiments and Results

To validate the presented deep learning decision fusion for the classification of urban remote sensing data, a series of comparative experiments are conducted on the widely used dataset of 2014 IEEE GRSS data fusion contest. It enables a challenging multiresolution and multisensor image analysis and data fusion opportunity; the visual data (Fig. 2) contain a series of color images associated with different strips, and the thermal hyperspectral image (Fig. 3) was acquired by the 84 spectral bands Hyper-Cam airborne sensor over Thetford Mines in Québec, Canada, with 874×751 pixels (spatial resolution of 1 m) and comes with a seven-class labeled ground truth map. In the above-described datasets, the training samples are randomly detached as 100 of each ground truth label, and the rest are used as the testing samples. In this section, a set of comparative experiments are carried out on the above-mentioned datasets to quantitatively investigate the significant advantages of the proposed classification frameworks over the conventional classifiers[69] containing decision tree (DT), discriminant analysis (DA), NB, k-nearest neighbor (KNN), and SVM. In the case of the conventional spectral-based classification of thermal hyperspectral image, we adopt eigenvalue (EV), hyperspectral signal subspace identification by minimum error (HS), and noise-whitened Harsanyi–Farrand–Chang (NH) techniques[36] as intrinsic dimension estimation to be employed by dimensionality reduction with principle components analysis (PCA).

The first experiment is carried out on the visible imagery of 2014 IEEE GRSS data fusion contest. To validate the proposed classification frameworks, two quality indices: overall accuracy (OA) and kappa coefficient are utilized to perform a comprehensive comparison. The quantitative evaluation results achieved by the various classifiers are shown in Table 2. It can be perceived that CNN gains the most accurate classification result (OA/kappa: 89.72/86.30) against the classic classifiers. Furthermore, the effectiveness of the presented classification techniques is evaluated via the visual inspection of the classification maps (Fig. 2). The second experiment is executed on the thermal hyperspectral imagery of 2014 IEEE GRSS data fusion contest. The implementation procedures are the same as that of the first experiment. As can be noted from Table 3, CNN obtains the highest classification accuracies (OA/kappa: 75.07/66.58). The classification maps of the multiple classifiers are shown in Fig. 3. The last experiment is conducted on a joint use of the above datasets. From the classification results in Tables 4 and 5, and by observing Fig. 4, it can be concluded that the presented CNN classification architecture provides again the best performance of classification (OA/kappa: 97.29/96.32).

The quantitative evaluation results obtained by the various classifiers are shown in Fig. 5. In addition, Fig. 6 summarizes the classification performance of the best classifiers with respect to the 2014 IEEE GRSS data fusion contest testing data. The overall results explain that the proposed deep learning frameworks perform better than the traditional classifiers in terms of metrics used. In this context, CNN provides 3.91/5.32%, 6.65/8.71%, 2.81/3.67%, and 5.52/7.37% enhancements for visual, thermal, combination, and context-aware OBP data, respectively, in terms of OA/kappa metrics. In contrast, the joint use of imaging systems improves performance of classification up to 7.57/10.02% and 22.22/29.74% with respect to the visible and thermal hyperspectral data, respectively. It is clearly obvious that the proposed method tends to be more robust and attains the highest classification results in terms of the classification quality index (Fig. 7).

An enormous number of hyperparameters for the above-mentioned classification frameworks can be automatically evaluated using the well-defined grid search procedure. In this case, Table 6 shows the grid search hyperparameters for the various classifiers.

In this section, we consider the dependencies of the presented deep learning frameworks applied for the experimental part, execution time analysis, and effect of model depth on

Table 2 The visual data classification accuracies.

No.	DT	DA	NB	KNN	SVM	SAE	CNN
1	0.65	0.68	0.60	0.79	0.86	0.90	0.88
2	0.82	0.76	**0.57**	0.80	0.81	0.85	0.91
3	0.84	0.88	**0.48**	0.87	0.93	0.90	0.95
4	0.74	0.66	**0.01**	0.68	0.61	0.74	0.83
5	0.94	0.97	0.98	0.95	0.96	**0.84**	0.93
6	0.71	0.85	0.58	0.83	0.87	0.78	0.93
7	**0.80**	0.95	0.88	0.93	0.95	0.93	0.95
OA	0.73	0.77	0.58	0.82	0.86	0.86	0.90
Kappa	0.66	0.70	0.47	0.76	0.81	0.81	0.86

Note: Bold values indicate outlier at the 5% level of significance.

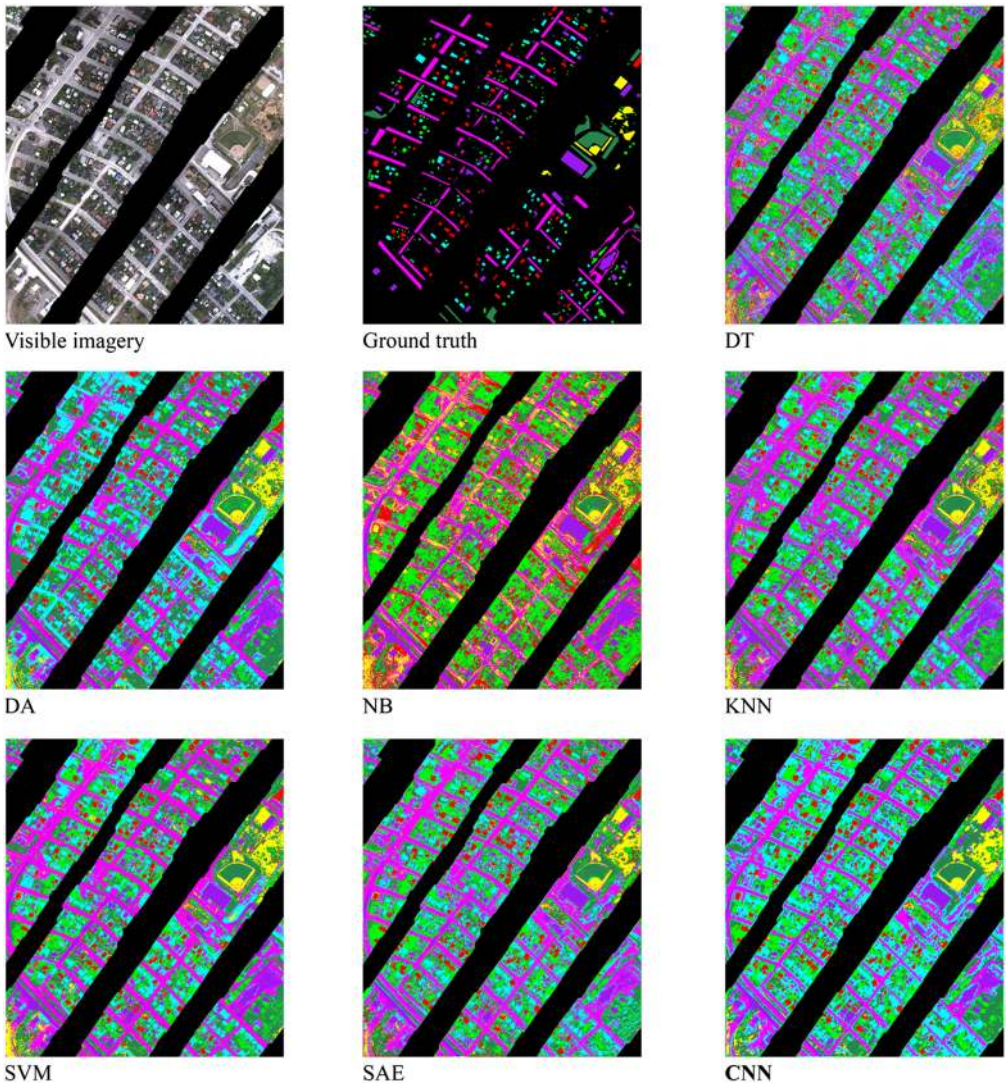

Fig. 2 Visible imagery, ground truth, and classification maps.

Table 3 The thermal hyperspectral data classification accuracies.

	PCA																
No.	EV	HS	NH	EV	HS	NH	EV	HS	NH	EV	HS	NH	EV	HS	NH	SAE	CNN
	DT			DA			NB			KNN			SVM				
1	0.73	0.82	0.88	0.80	0.91	0.91	0.71	0.91	0.90	0.69	0.92	0.90	0.70	0.90	0.89	0.93	0.95
2	0.38	0.40	0.29	**0.08**	0.25	0.40	0.34	0.42	0.42	0.46	0.50	0.42	0.31	0.36	0.37	0.33	0.54
3	0.47	0.40	0.41	0.34	0.48	0.42	0.30	0.56	0.52	0.43	0.55	0.59	0.16	0.60	0.48	0.46	0.54
4	0.13	0.43	0.39	0.27	0.44	0.49	0.35	0.37	0.43	0.18	0.40	0.32	0.51	0.40	0.49	0.47	0.53
5	0.30	0.44	0.46	0.12	0.32	0.34	0.18	0.22	0.23	0.28	0.43	0.33	0.01	0.41	0.28	0.50	0.67
6	0.32	0.37	0.48	0.65	0.59	0.45	0.58	0.56	0.56	0.23	0.37	0.37	0.60	0.56	0.55	0.52	0.55
7	0.46	0.61	0.52	**0.01**	0.45	0.48	0.44	0.44	0.45	0.38	0.63	0.59	0.37	0.66	0.57	0.66	0.77
OA	0.51	0.61	0.64	0.53	0.66	0.65	0.53	0.65	0.65	0.48	0.67	0.64	0.51	0.68	0.65	0.69	0.75
Kappa	0.37	0.48	0.51	0.37	0.54	0.53	0.39	0.54	0.54	0.34	0.56	0.52	0.37	0.58	0.54	0.58	0.67

Note: Number of intrinsic dimensionality: EV = 1, HS = 5, and NH = 13. Bold values indicate outlier at the 5% level of significance.

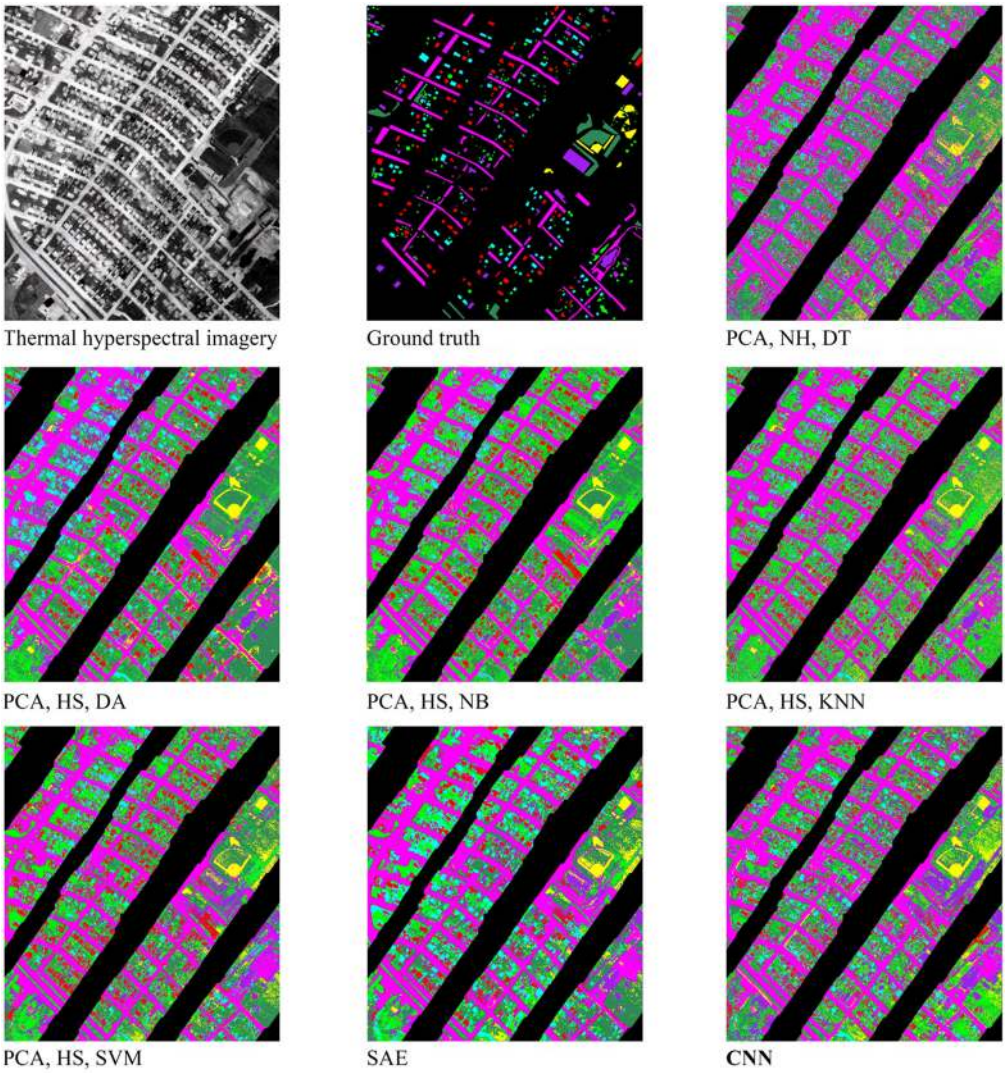

Fig. 3 Thermal hyperspectral imagery, ground truth, and classification maps.

Table 4 NB combination classification accuracies.

	PCA																
No.	EV	HS	NH	EV	HS	NH	EV	HS	NH	EV	HS	NH	EV	HS	NH	SAE	CNN
		DT			DA			NB			KNN			SVM			
1	0.70	0.82	0.88	0.79	0.93	0.93	0.75	0.93	0.92	0.78	0.94	0.92	0.87	0.91	0.94	0.93	0.89
2	0.81	0.76	0.82	0.74	0.79	0.76	0.63	0.52	0.54	0.74	0.75	0.76	0.77	0.77	0.76	0.85	0.91
3	0.84	0.83	0.87	0.90	0.93	0.92	0.69	0.76	0.63	0.92	0.94	0.92	0.95	0.96	0.95	0.90	0.95
4	0.73	0.81	0.84	0.71	0.87	0.87	0.37	0.45	0.47	0.69	0.85	0.89	0.72	0.88	0.87	0.74	0.93
5	0.93	0.93	0.93	0.94	0.96	0.96	0.95	0.94	0.97	0.93	0.95	0.94	0.96	0.94	0.96	**0.82**	0.92
6	0.72	0.73	0.71	0.81	0.77	0.84	0.53	0.64	0.65	0.83	0.84	0.84	0.87	0.88	0.87	0.78	0.93
7	0.85	0.90	0.88	0.93	0.95	0.95	0.86	0.90	0.89	0.82	0.94	0.93	0.95	0.95	0.96	0.93	0.95
OA	0.76	0.82	0.85	0.81	0.89	0.90	0.69	0.80	0.79	0.81	0.90	0.90	0.87	0.90	0.91	0.87	0.91
Kappa	0.69	0.76	0.80	0.76	0.86	0.87	0.61	0.73	0.72	0.75	0.87	0.86	0.83	0.87	0.89	0.82	0.89

Note: Number of intrinsic dimensionality: EV = 1, HS = 5, and NH = 13. Bold values indicate outlier at the 5% level of significance.

Table 5 BKS combination classification accuracies.

	PCA																	
No.	EV	HS	NH	EV	HS	NH	EV	HS	NH	EV	HS	NH	EV	HS	NH	SAE	CNN	OBP
		DT			DA			NB			KNN			SVM				
1	0.71	0.85	0.88	0.64	0.95	0.93	0.69	0.95	0.92	0.66	0.94	0.90	0.71	0.91	0.94	0.98	0.98	1.00
2	0.81	0.76	0.82	0.73	0.80	0.75	0.63	0.65	0.67	0.74	0.75	0.77	0.81	0.77	0.81	0.84	0.89	0.89
3	0.82	0.83	0.85	0.86	0.93	0.90	0.74	0.75	0.73	0.89	0.94	0.94	0.95	0.92	0.94	0.89	0.94	0.97
4	0.75	0.83	0.84	0.84	0.82	0.88	0.33	0.45	0.39	0.79	0.88	0.87	0.85	0.90	0.89	0.72	0.87	0.96
5	0.93	0.93	0.93	0.94	0.96	0.96	0.95	0.94	0.97	0.93	0.88	0.95	0.95	0.89	0.96	**0.82**	0.92	0.96
6	0.72	0.72	0.70	0.85	0.77	0.84	0.45	0.45	0.45	0.83	0.83	0.82	0.87	0.85	0.87	0.78	0.93	0.95
7	0.85	0.89	0.88	0.96	0.94	0.95	0.89	0.88	0.89	0.84	0.94	0.93	0.95	0.95	0.96	0.92	0.95	1.00
OA	0.76	0.83	0.85	0.76	0.90	0.90	0.66	0.79	0.77	0.76	0.90	0.89	0.81	0.89	0.92	0.89	0.95	0.97
Kappa	0.69	0.78	0.80	0.70	0.87	0.87	0.56	0.72	0.70	0.69	0.87	0.85	0.76	0.86	0.89	0.85	0.93	0.96

Note: Number of intrinsic dimensionality: EV = 1, HS = 5, and NH = 13. Bold values indicate outlier at the 5% level of significance.

classification of the 2014 IEEE GRSS data fusion contest. In this context, the presented deep learning architectures are first described in Table 7.

The execution time of the deep learning architecture consists of training and testing times; training time demonstrates the time utilization of the learning filters, classification layers, and fine-tuning the deep feature learning framework. Figure 8 exhibits how the training time changes with variation in model neurons and iteration epoch parameters. It can be observed that the training time gradually increases by the extension of the number of layer neurons and iteration epochs.

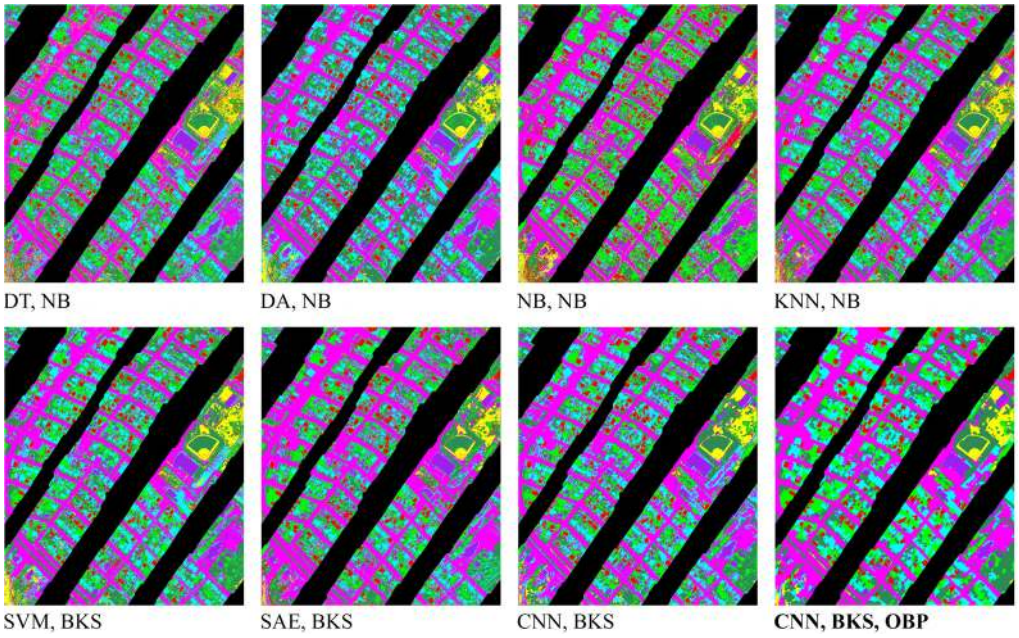

Fig. 4 Combination classification maps.

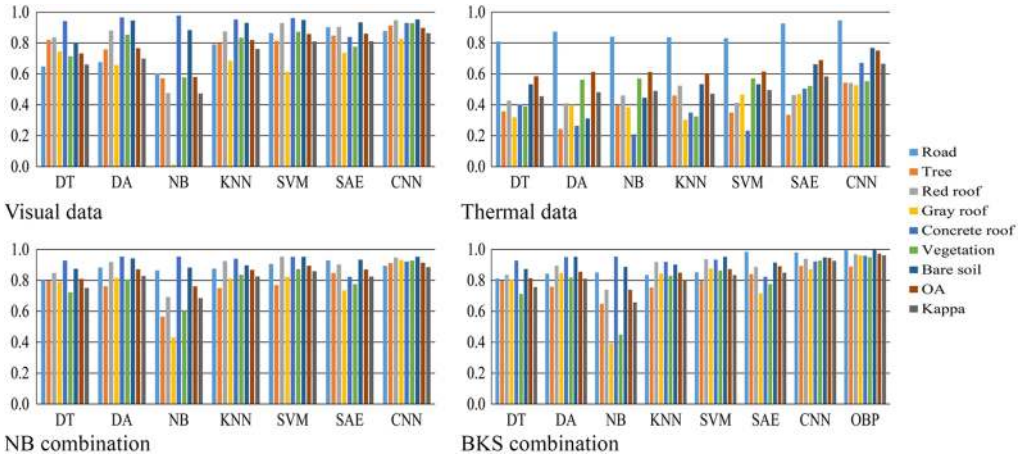

Fig. 5 The quantitative evaluation results obtained by the various classifiers.

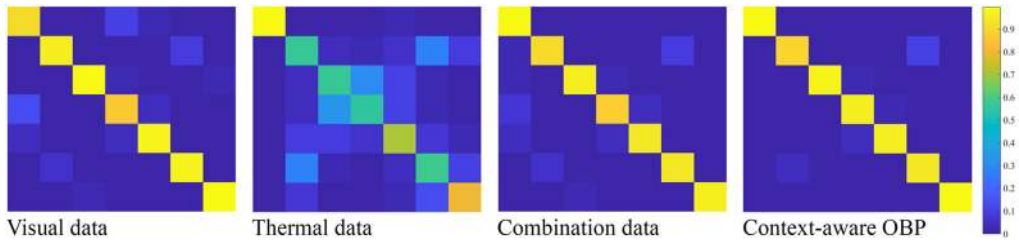

Fig. 6 The confusion matrix of the best classifiers.

Model depth additionally plays a considerable role in the performance of classification because it can enhance the feature representation quality of the input data. Mainly, the higher model depths tend to exploit more invariant and abstract features of the raw data. Within this context, a series of comparative experiments are carried out to assess how the depth parameter defines a significant role in the classification performance (Table 8). It can be noted

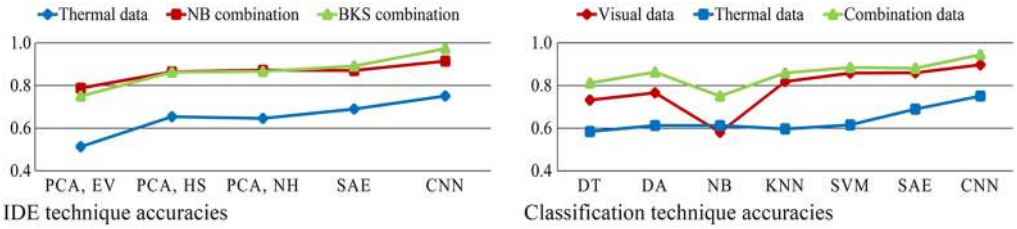

Fig. 7 Average OA of the classification results.

Table 6 The hyperparameters of the various classifiers on the dataset of 2014 IEEE GRSS data fusion contest.

			The dataset of 2014 IEEE GRSS data fusion contest	
Classifier	Hyperparameters		Visual data	Thermal data
DT	Minimum number of leaf node observations		1	12
DA	Linear coefficient threshold		0	0
	Amount of regularization		0	1
NB	Data distributions		Kernel	Normal
	Kernel smoothing window width		0.015	—
KNN	Distance metric		Euclidean	Euclidean
	Number of nearest neighbors to find		5	5
SVM	Coding design		onevsone	onevsone
	Kernel function		Gaussian	Gaussian
	Box constraint		100	100
	Kernel scale parameter		1	0.25
SAE	Size of input data		[5 5 3]	[1 83] – [3 3 3]
	Size of hidden representation of the autoencoder		15	25
	Maximum number of training epochs		1000	1000
	Desired proportion of training examples a neuron reacts to		0.50	0.50
CNN	Size of input data		[7 7 3]	[1 83] – [5 5 5]
	Height and width of filters		[4 4] – [2 2]	[1 7] – [1 6] – [1 7]
	Number of filters		16 – 32	16 – 24 – 48
	Height and width of pooling region		[2 2] – [1 1]	[1 2] – [1 2] – [1 2]
	Step size for traversing the input		[1 1] – [1 1]	[1 2] – [1 2] – [1 2]
	Initial learning rate		0.01	0.01
	Maximum number of training epochs		1000	1000
OBP	Multiresolution segmentation	Scale parameter	15	
		Shape	0.25	
		Compactness	0.50	
	Threshold of length to width		4.0	
	Threshold of area pixels		400	

Table 7 The proposed deep learning architectures.

	SAE						CNN					
								Number of filters				
								C2	C4	C6		
Dataset	I1	AE2	AE3	AE4	F5	O6	I1	S3	S5	S7	F8	O9
Visual data	$5\times5\times3$	15	15	15	Fully connected layer	1×7	$7\times7\times3$	16 4×4 2×2	32 2×2 1×1	–	Fully connected layer	1×7
Thermal data	1×110	25	25	25	Fully connected layer	1×7	1×208	16 1×7 1×2	24 1×6 1×2	48 1×7 1×2	Fully connected layer	1×7

Note: I, input layer; AE, autoencoder layers; C, convolution layers; S, pooling layers; F, fully connected layer; O, output layer.

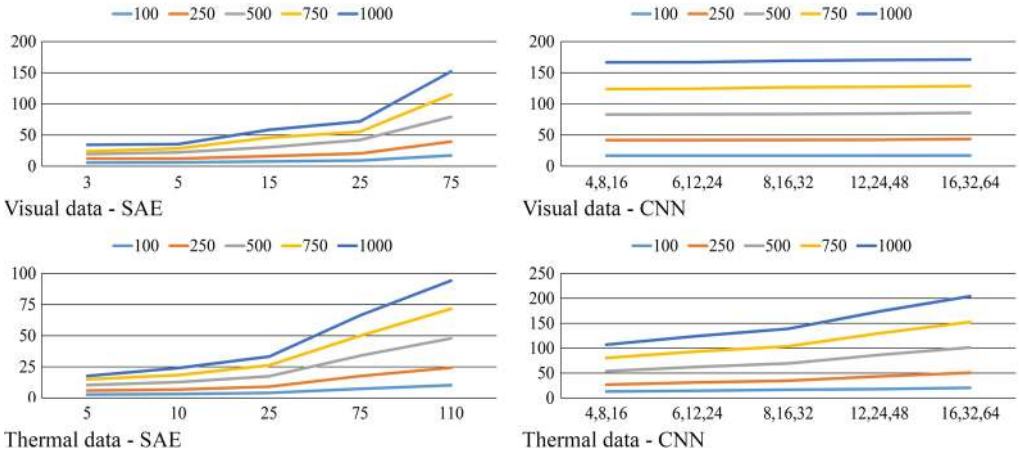

Fig. 8 Comparison of training time with the variation of model parameters.

Table 8 Comparison of OA for different model depths.

		SAE		CNN	
Depth analysis		Visual data	Thermal data	Visual data	Thermal data
Layer 1	OA	81.86	64.01	89.05	73.24
	Training time	31.82	19.53	160.82	215.26
	Testing time	1.25	1.21	45.91	23.36
Layer 2	OA	84.04	67.02	89.72	73.66
	Training time	45.81	26.35	172.23	291.83
	Testing time	1.29	1.39	46.03	33.66
Layer 3	OA	85.97	68.95	88.30	75.07
	Training time	59.74	33.99	190.50	368.15
	Testing time	1.49	1.65	50.99	45.48

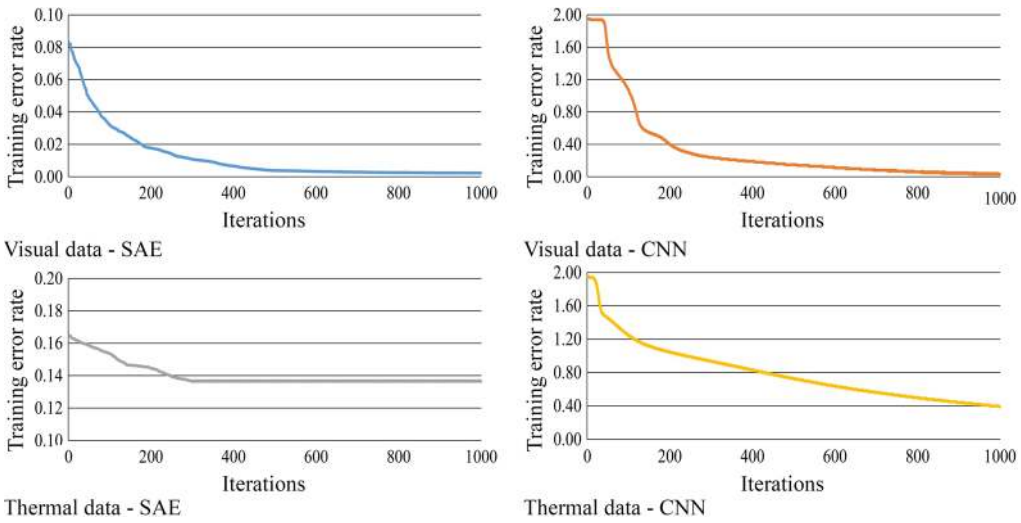

Fig. 9 Convergence curves of the training samples.

that the performance of classification enhances the expansion of the model depth parameter; simultaneously, the execution time gradually increases by the extension of the number of layers and iteration epochs. In addition, Fig. 9 shows the convergence curves of the training samples.

The obtained results prove the significant advantages of the proposed spectral–spatial deep learning frameworks over the conventional spectral-based classification methods.

4 Conclusion

In this paper, joint spectral–spatial information is extracted in deep learning frameworks for classification of urban remote sensing data. A series of comparative experiments indicate that the spectral–spatial feature descriptors enhance the performance of classification compared with the conventional spectral-based classifiers. Based on consistency over the widely used dataset of 2014 IEEE GRSS data fusion contest, the presented frameworks provide statistically higher classification accuracy and appear to be more robust than the traditional classifiers. Execution time and effect of model depth on the above-mentioned dataset were evaluated by a set of experiments. We suggest applying a deep learning model to achieve higher classification accuracy and consume the least amount of execution time. In our future work, we will consider how to apply pretrained networks to save huge efforts required to retrain the deep learning architecture.

Acknowledgments

The authors would like to thank Telops Inc. (Québec, Canada) for acquiring and providing the data used in this study, the IEEE GRSS Image Analysis and Data Fusion Technical Committee and Dr. M. Shimoni (Signal and Image Centre, Royal Military Academy, Belgium) for organizing the 2014 Data Fusion Contest, the Centre de Recherche Public Gabriel Lippmann (CRPGL, Luxembourg) and Dr. M. Schlerf (CRPGL) for their contribution of the HyperCam LWIR sensor, and Dr. M. De Martino (University of Genoa, Italy) for her contribution in data preparation.

References

1. S. Li et al., "An effective feature selection method for hyperspectral image classification based on genetic algorithm and support vector machine," *Knowl.-Based Syst.* **24**(1), 40–48 (2011).

2. P. Du et al., "Information fusion techniques for change detection from multi-temporal remote sensing images," *Inf. Fusion* **14**(1), 19–27 (2013).
3. B. Bigdeli, F. Samadzadegan, and P. Reinartz, "A multiple SVM system for classification of hyperspectral remote sensing data," *J. Indian Soc. Remote Sens.* **41**(4), 763–776 (2013).
4. B. Bigdeli, F. Samadzadegan, and P. Reinartz, "A decision fusion method based on multiple support vector machine system for fusion of hyperspectral and LIDAR data," *Int. J. Image Data Fusion* **5**(3), 196–209 (2014).
5. X. Lu et al., "Synergetic classification of long-wave infrared hyperspectral and visible images," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **8**(7), 3546–3557 (2015).
6. L. Wald, "Some terms of reference in data fusion," *IEEE Trans. Geosci. Remote Sens.* **37**(3), 1190–1193 (1999).
7. C. Thomas et al., "Synthesis of multispectral images to high spatial resolution: a critical review of fusion methods based on remote sensing physics," *IEEE Trans. Geosci. Remote Sens.* **46**(5), 1301–1312 (2008).
8. A. Plaza et al., "Recent advances in techniques for hyperspectral image processing," *Remote Sens. Environ.* **113**(1), S110–S122 (2009).
9. M. Fauvel et al., "Advances in spectral-spatial classification of hyperspectral images," *Proc. IEEE* **101**(3), 652–675 (2013).
10. G. Camps-Valls et al., "Advances in hyperspectral image classification: earth monitoring with statistical learning methods," *IEEE Signal Process Mag.* **31**(1), 45–54 (2014).
11. J. Tian and P. Reinartz, "Multitemporal 3D change detection in urban areas using stereo information from different sensors," in *Int. Symp. on Image and Data Fusion (ISIDF 2011)*, IEEE, pp. 1–4 (2011).
12. L. Bruzzone and F. Bovolo, "A novel framework for the design of change-detection systems for very-high-resolution remote sensing images," *Proc. IEEE* **101**(3), 609–630 (2013).
13. D. Espinoza-Molina and M. Datcu, "Earth-observation image retrieval based on content, semantics, and metadata," *IEEE Trans. Geosci. Remote Sens.* **51**(11), 5145–5159 (2013).
14. P. Blanchart et al., "Pattern retrieval in large image databases using multiscale coarse-to-fine cascaded active learning," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **7**(4), 1127–1141 (2014).
15. C. Wemmert et al., "Multiresolution remote sensing image clustering," *IEEE Geosci. Remote Sens. Lett.* **6**(3), 533–537 (2009).
16. A. Voisin et al., "Supervised classification of multisensor and multiresolution remote sensing images with a hierarchical copula-based approach," *IEEE Trans. Geosci. Remote Sens.* **52**(6), 3346–3358 (2014).
17. L. Bruzzone and M. Marconcini, "Toward the automatic updating of land-cover maps by a domain-adaptation SVM classifier and a circular validation strategy," *IEEE Trans. Geosci. Remote Sens.* **47**(4), 1108–1122 (2009).
18. C. Persello and L. Bruzzone, "Active learning for domain adaptation in the supervised classification of remote sensing images," *IEEE Trans. Geosci. Remote Sens.* **50**(11), 4468–4483 (2012).
19. D. Tuia et al., "Semisupervised manifold alignment of multimodal remote sensing images," *IEEE Trans. Geosci. Remote Sens.* **52**(12), 7708–7720 (2014).
20. D. Tuia et al., "A survey of active learning algorithms for supervised remote sensing image classification," *IEEE J. Sel. Top. Signal Process.* **5**(3), 606–617 (2011).
21. M. M. Crawford, D. Tuia, and H. L. Yang, "Active learning: any value for classification of remotely sensed data?" *Proc. IEEE* **101**(3), 593–608 (2013).
22. G. Moser, S. B. Serpico, and J. A. Benediktsson, "Land-cover mapping by Markov modeling of spatial–contextual information in very-high-resolution remote sensing images," *Proc. IEEE* **101**(3), 631–651 (2013).
23. J. Li et al., "Urban classification by the fusion of thermal infrared hyperspectral and visible data," *Photogramm. Eng. Remote Sens.* **81**(12), 901–911 (2015).
24. W. Liao et al., "Processing of multiresolution thermal hyperspectral and digital color data: outcome of the 2014 IEEE GRSS data fusion contest," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **8**(6), 2984–2996 (2015).

25. M. Eslami and A. Mohammadzadeh, "Developing a spectral-based strategy for urban object detection from airborne hyperspectral TIR and visible data," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **9**(5), 1808–1816 (2016).
26. M. Eslami and A. Mohammadzadeh, "A novel method for urban land cover mapping based on new vegetation indices and texture-spectral information from fused visible and hyperspectral thermal infrared airborne data," *Eur. J. Remote Sens.* **50**(1), 320–331 (2017).
27. F. Samadzadegan, H. Hasani, and P. Reinartz, "Toward optimum fusion of thermal hyperspectral and visible images in classification of urban area," *Photogramm. Eng. Remote Sens.* **83**(4), 269–280 (2017).
28. G. Abdi, F. Samadzadegan, and P. Reinartz, "A decision-based multi-sensor classification system using thermal hyperspectral and visible data in urban area," *Eur. J. Remote Sens.* **50**(1), 414–427 (2017).
29. F. Samadzadegan, H. Hasani, and T. Schenk, "Simultaneous feature selection and SVM parameter determination in classification of hyperspectral imagery using ant colony optimization," *Can. J. Remote Sens.* **38**(2), 139–156 (2014).
30. P. Pahlavani and B. Bigdeli, "A mutual information-Dempster–Shafer based decision ensemble system for land cover classification of hyperspectral data," *Front. Earth Sci.* **11**(4), 774–783 (2017).
31. G. R. Hunt, "Spectral signatures of particulate minerals in the visible and near infrared," *Geophysics* **42**(3), 501–513 (1977).
32. B. D. Saksena, "Infra-red absorption studies of some silicate structures," *Trans. Faraday Soc.* **57**, 242–258 (1961).
33. O. Sandus, "A review of emission polarization," *Appl. Opt.* **4**(12), 1634–1642 (1965).
34. D. C. Bertilone, "Stokes parameters and partial polarization of far-field radiation emitted by hot bodies," *J. Opt. Soc. Am. A* **11**(8), 2298–2304 (1994).
35. K. Segl et al., "Fusion of spectral and shape features for identification of urban surface cover types using reflective and thermal hyperspectral data," *ISPRS J. Photogramm. Remote Sens.* **58**(1-2), 99–112 (2003).
36. C.-I. Chang, *Hyperspectral Data Processing: Algorithm Design and Analysis*, John Wiley & Sons, Hoboken, New Jersey (2013).
37. C.-H. Li et al., "A spatial–contextual support vector machine for remotely sensed image classification," *IEEE Trans. Geosci. Remote Sens.* **50**(3), 784–799 (2012).
38. L. Fang et al., "Spectral–spatial hyperspectral image classification via multiscale adaptive sparse representation," *IEEE Trans. Geosci. Remote Sens.* **52**(12), 7738–7749 (2014).
39. D. Akbari et al., "Mapping urban land cover based on spatial-spectral classification of hyperspectral remote-sensing data," *Int. J. Remote Sens.* **37**(2), 440–454 (2016).
40. L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: a technical tutorial on the state of the art," *IEEE Geosci. Remote Sens. Mag.* **4**(2), 22–40 (2016).
41. X. Chen et al., "Vehicle detection in satellite images by hybrid deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.* **11**(10), 1797–1801 (2014).
42. Y. Chen et al., "Deep learning-based classification of hyperspectral data," *IEEE J. Sel. Top. Appl. Earth Obs. Remote sens.* **7**(6), 2094–2107 (2014).
43. C. Tao et al., "Unsupervised spectral–spatial feature learning with stacked sparse autoencoder for hyperspectral imagery classification," *IEEE Geosci. Remote Sens. Lett.* **12**(12), 2438–2442 (2015).
44. J. Yue et al., "Spectral–spatial classification of hyperspectral images using deep convolutional neural networks," *Remote Sens. Lett.* **6**(6), 468–477 (2015).
45. K. Makantasis et al., "Deep learning-based man-made object detection from hyperspectral data," in *Int. Symp. on Visual Computing*, pp. 717–727 (2015).
46. H. Liang and Q. Li, "Hyperspectral imagery classification using sparse representations of convolutional neural network features," *Remote Sens.* **8**(2), 99 (2016).
47. M. Längkvist et al., "Classification and segmentation of satellite orthoimagery using convolutional neural networks," *Remote Sens.* **8**(4), 329 (2016).
48. C. Zhao et al., "Spectral–spatial classification of hyperspectral imagery based on stacked sparse autoencoder and random forest," *Eur. J. Remote Sens.* **50**(1), 47–63 (2017).

49. L. Wang et al., "Spectral–spatial multi-feature-based deep learning for hyperspectral remote sensing image classification," *Soft Comput.* **21**(1), 213–221 (2017).
50. S. Yu, S. Jia, and C. Xu, "Convolutional neural networks for hyperspectral image classification," *Neurocomputing* **219**, 88–98 (2017).
51. P. Ghamisi et al., "Advanced spectral classifiers for hyperspectral images: a review," *IEEE Geosci. Remote Sens. Mag.* **5**(1), 8–32 (2017).
52. Y. Li, H. Zhang, and Q. Shen, "Spectral-spatial classification of hyperspectral imagery with 3D convolutional neural network," *Remote Sens.* **9**(1), 67 (2017).
53. G. Abdi, F. Samadzadegan, and P. Reinartz, "Spectral-spatial feature learning for hyperspectral imagery classification using deep stacked sparse autoencoder," *J. Appl. Remote Sens.* **11**(4), 042604 (2017).
54. T. Serre et al., "A quantitative theory of immediate visual recognition," *Progr. Brain Res* **165**, 33–56 (2007).
55. Y. Freund and D. Haussler, "Unsupervised learning of distributions of binary vectors using two layer networks," in *Advances in Neural Information Processing Systems*, pp. 912–919 (1994).
56. G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.* **18**(7), 1527–1554 (2006).
57. D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. **1**, pp. 318–362, MIT Press, Cambridge, Massachusetts (1986).
58. P. Vincent et al., "Extracting and composing robust features with denoising autoencoders," in *Proc. of the 25th Int. Conf. on Machine Learning*, pp. 1096–1103 (2008).
59. H. Lee et al., "Efficient sparse coding algorithms," in *Advances in Neural Information Processing Systems*, Vol. **19**, p. 801 (2007).
60. Y. LeCun et al., "Gradient-based learning applied to document recognition," *Proc. IEEE* **86**(11), 2278–2324 (1998).
61. Y. Bengio et al., "Learning deep architectures for AI," *Found. Trends® Mach. Learn.* **2**(1), 1–127 (2009).
62. Y. Bengio, A. Courville, and P. Vincent, "Representation learning: a review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(8), 1798–1828 (2013).
63. G. E. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science* **313**(5786), 504–507 (2006).
64. C. C. Tan and C. Eswaran, "Reconstruction of handwritten digit images using autoencoder neural networks," in *Canadian Conf. on Electrical and Computer Engineering (CCECE 2008)*, pp. 465–470 (2008).
65. X. Zhang et al., "Fusing heterogeneous features from stacked sparse autoencoder for histopathological image analysis," *IEEE J. Biomed. Health Inf.* **20**(5), 1377–1383 (2016).
66. D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Cognit. Model.* **5**(1), 533–536 (1988).
67. D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Math. Program.* **45**, 503–528 (1989).
68. Y. S. Huang and C. Y. Suen, "A method of combining multiple experts for the recognition of unconstrained handwritten numerals," *IEEE Trans. Pattern Anal. Mach. Intell.* **17**(1), 90–94 (1995).
69. L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*, 2nd ed., John Wiley & Sons, Hoboken, New Jersey (2014).
70. B. Bigdeli and P. Pahlavani, "High resolution multisensor fusion of SAR, optical and LiDAR data based on crisp vs. fuzzy and feature vs. decision ensemble systems," *Int. J. Appl. Earth Obs. Geoinf.* **52**, 126–136 (2016).
71. M. Baatz and S. Arno, "Multiresolution segmentation: an optimization approach for high quality multiscale image segmentation," in *Angewandte Geographische Informationsverarbeitung XII*, Vol. **58**, pp. 12–23 (2000).

Ghasem Abdi received his BSc degree in geomatics engineering and his MSc degree in photogrammetry engineering from the University of Tehran, Tehran, Iran, in 2009 and 2012,

respectively. Currently, he is working toward his PhD in photogrammetry engineering at the University of Tehran, Tehran, Iran. His research interests include computer vision and pattern recognition, deep learning and machine vision, and image processing.

Farhad Samadzadegan received his PhD in photogrammetry engineering from the University of Tehran, Tehran, Iran, in 2001. Currently, he is working as a full professor in the faculty of surveying and geospatial engineering at the University of Tehran, Tehran, Iran. He has more than 15 years of experience in designing and developing digital photogrammetric and remote sensing software and systems.

Peter Reinartz received his PhD in civil engineering from the University of Hannover, Hannover, Germany, in 1989. He is the head of the Department of Photogrammetry and Image Analysis, German Aerospace Centre (DLR), Remote Sensing Technology Institute, Wessling, Germany, and holds a professorship for geomatics at the University of Osnabrueck, Osnabrueck, Germany. He has more than 30 years of experience in image processing and remote sensing and over 400 papers in these fields.