Pattern Recognition 61 (2017) 379–393

Contents lists available at ScienceDirect

Pattern Recognition

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

Radar HRRP target recognition with deep networks

Bo Feng [a,b], Bo Chen [a,b,*], Hongwei Liu [a,b]

[a] National Lab of Radar Signal Processing, Xidian University, Xi'an 710071, China
[b] Collaborative Innovation Center of Information Sensing and Understanding at Xidian University, Xi'an 710071, China

ARTICLE INFO

Article history:
Received 8 August 2014
Received in revised form
12 June 2016
Accepted 15 August 2016
Available online 16 August 2016

Keywords:
Radar automatic target recognition (RATR)
High-resolution range profile (HRRP)
Deep networks
Stacked Corrective Autoencoders (SCAE)

ABSTRACT

Feature extraction is the key technique for radar automatic target recognition (RATR) based on high-resolution range profile (HRRP). Traditional feature extraction algorithms usually utilize shallow architectures, which result in the limited capability to characterize HRRP data and restrict the generalization performance for RATR. Aiming at those issues, in this paper deep networks are built up for HRRP target recognition by adopting multi-layered nonlinear networks for feature learning. To learn the stable structure and correlation of targets from unlabeled data, a deep network called Stacked Corrective Autoencoders (SCAE) is further proposed via taking the advantage of the HRRP's properties. As an extension of deep autoencoders, SCAE is stacked by a series of Corrective Autoencoders (CAE) and employs the average profile of each HRRP frame as the correction term. The covariance matrix of each HRRP frame is considered for establishing an effective loss function under the Mahalanobis distance criterion. We use the measured HRRP data to show the effectiveness of our methods. Furthermore, we demonstrate that with the proper optimization procedure, our model is also effective even with a moderately incomplete training set.

1. Introduction

A high-resolution range profile (HRRP) is the amplitude of the coherent summations of the complex time returns from target scatterers in each range cell, which represents the projection of the complex returned echoes from the target scattering centers onto the radar line-of-sight (LOS). It contains abundant target structure signatures, such as target size, scatterer distribution, etc., thereby radar HRRP target recognition has received intensive attention from the radar automatic target recognition (RATR) community [1–17].

In many applications of machine learning, such as HRRP-based RATR, feature extraction is fundamental and crucial for different systems. The performance of these applications is heavily dependent on learned data features on which they are applied [18]. For that reason, many researchers [3–12] pay a lot attention to explore various feature extraction methods. Some researchers [5] investigate the bispectra feature extracted from real HRRP data and Du et al. [6] analyze the recognition performance of power spectrum, FFT-magnitude feature [3,7], and various high-order spectra features. In [8], cepstrum- and bicoherence-based features from a sequence of radar returned echoes are proposed and analyzed. Such engineered features are useful but rely on researchers' experience and skill. If we do not have sufficient prior knowledge for the applications, those features would be brittle and incomplete. Feng et al. [7] utilize dictionary learning to represent the favor sparse overcomplete features of HRRP data, and achieve better generalization performance. In [2], the PCA-based feature subspace is constructed to minimize reconstruction error for RATR. Some super-resolution algorithms are also employed to extract the location information of predominant scatterers from real HRRP data as features [11,12]. Those methods are based on particular parameter models and work well in practice, but all of them are shallow architectures. Recently, many researchers [18–29] suggest that non-linear, deep networks can achieve superior performance in various real world tasks.

In the area of unsupervised feature learning, deep networks such as Deep Belief Networks (DBN) [20] and Stacked Denoising Autoencoders (SDAE) [21] have been widely applied to various machine learning tasks. Those models are distinguished from the conventional multi-layer perceptron (MLP) and other shallow models by their large number of layers of features and their adaptation by a layer-wise unsupervised greedy learning method [18]. Deep networks learn to identify and disentangle the underlying explanatory factors hidden in the observed milieu of low-level sensory data. They are expected to perform well in machine learning problems because of the similar hierarchical structure to human visual cortex [18,24].

In this paper, we build up a recognition procedure based on deep networks. To our best knowledge, we are the first to use deep

* Corresponding author at: National Laboratory of Radar Signal Processing and Collaborative Innovation Center of Information Sensing and Understanding at Xidian University, Xi'an 710071, China.
E-mail address: bchen@mail.xidian.edu.cn (B. Chen).

http://dx.doi.org/10.1016/j.patcog.2016.08.012
0031-3203/

networks in HRRP RATR area. Several issues should be considered when "deep" architecture is applied to HRRP-based RATR. Such as the reduction of time-shift, target-aspect and amplitude-scale sensitivity need to be done by preprocessing or be embedded in the deep networks. We then test two conventional training schemes, i.e., DBN and SDAE. The experimental results show these deep networks have better recognition performance than several existing feature extraction methods, like Principal component analysis (PCA) [2], K-SVD [9], Linear discriminant analysis (LDA) [30], etc. However, such deep networks are designed without the consideration of the HRRP's properties. Due to the 'speckle effect' [2,8,15], HRRPs may fluctuate significantly even with slight aspect variations, i.e., a slight change in elevation or azimuth aspect may provoke considerable variation in the intensity of peaks (amplitude fluctuation property), while the location of the peaks remain unchanged (structure similarity). These properties, which are ignored in the models, can affect the final recognition performance. Consequently, we first analyze these and then based on it we propose a novel deep architecture named Stacked Corrective Autoencoders (SCAE) for HRRP target recognition.

Moreover, in the non-cooperative circumstance, the amount of data under test usually is huge, but the training data is limited or even the HRRP data from some aspects may be missing or non-observed. Such cases will lead to overfitting and restrict the generalization performance of the model. Detailed experiments have been also done to examine and analyze the generalization performance of SCAE under those cases.

The reminder of the paper is organized as follows. A brief description of HRRP and deep networks is provided in Section 2. Then we propose a novel deep architecture, i.e., SCAE for radar HRRP target recognition in Section 3. The parameter setting and detailed experimental results based on measured HRRP data are provided in Section 4, followed by conclusions in Section 5.

2. Preliminaries

In this section, we will briefly review the concepts of HRRP and deep networks. A general description of HRRP is included to present its properties at first. Since our models are built upon networks, we then introduce some basic notations and give preliminaries about Restricted Boltzmann Machine (RBM) [20,22], Autoencoder (AE) [21] as well as their corresponding extensions, e.g. Gaussian–Bernoulli RBM (GRBM) [29] and Denoising Autoencoder (DAE) [21,25], all of which are the fundamental building blocks of deep networks.

2.1. Description of HRRP

HRRP can be considered to be the amplitude of the coherent summations of the complex time returns from target scatterers in each range cell, which represents the reflected radar intensity versus range along the target extent, as shown in Fig. 1.

HRR radar operates in microwave frequency band. Generally, the size of the targets or their components is much larger than the wavelength of the radar. For complex targets such as an aircraft, HRR radar can effectively divide the object into many range "cells". The radar signatures from scattering centers within the same range cell will be coherently summed into a single signature for that rang cell. According to the literatures [2,3], if the radar transmitted signal is $s(t)e^{j2\pi f_c t}$, the nth complex returned echo in the dth range cell ($d = 1, 2, \ldots, D$) in baseband can be approximated as

$$\tilde{x}_d(t, n) \approx s(t) \sum_{i=1}^{L_d} \sigma_{di} e^{-j\{(4\pi/\lambda)[R(n)+\Delta r_{di}(n)]\}} \tag{1}$$

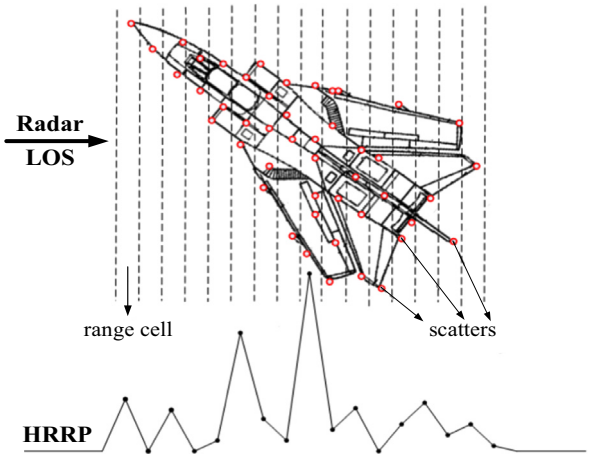

Fig. 1. Illustration of an HRRP sample from an aircraft target. The circles on the aircraft represent some target scatterers. This figure is cited from [16].

where $s(t)$ is the complex envelop which approximates to be unvaried with the radial displacements for all scatterers in one range cell. f_c is the radar signal carrier frequency and λ denotes the wavelength of the HRR radar. L_d represents the number of target scatterers in the dth range cell. σ_{di} is the intensity of the ith scatterer in the dth range cell. $R(n)$ is the radial distance between the target reference center in the nth returned echo and the radar. $\Delta r_{di}(n)$ is the radial displacement of the ith scatterer of the dth range cell in the nth returned echo. When $s(t)$ is a rectangular pulse signal with unit intensity, it usually be omitted. After eliminating the initial phase of the nth returned echo $e^{-j(4\pi/\lambda)R(n)}$, the nth HRRP can be defined as

$$\begin{aligned}\mathbf{x}(n) &\triangleq \left[x_1(n), x_2(n), \ldots, x_D(n)\right] \\ &= \left[\left|\sum_{i=1}^{L_1} \sigma_{1i} e^{j\phi_{1i}(n)}\right|, \left|\sum_{i=1}^{L_2} \sigma_{2i} e^{j\phi_{2i}(n)}\right|, \ldots, \left|\sum_{i=1}^{L_D} \sigma_{1i} e^{j\phi_{Di}(n)}\right|\right]\end{aligned} \tag{2}$$

where $\phi_{di}(n) = -(4\pi/\lambda)\Delta r_{di}(n)$.

The power of the dth range cell $x_d(n)$ in $\mathbf{x}(n)$ is

$$\begin{aligned}x_d^2(n) = &\sum_{i=1}^{L_d} \sigma_{di}^2 e^{j\phi_{di}(n)} \\ &+ 2\sum_{i=2}^{L_d} \sum_{k=1}^{i-1} \sigma_{di}\sigma_{dk} \cos\left[\phi_{di}(n) - \phi_{dk}(n)\right]\end{aligned} \tag{3}$$

Due to the scattering center amplitude and phase variations and interactions between returns from individual scattering

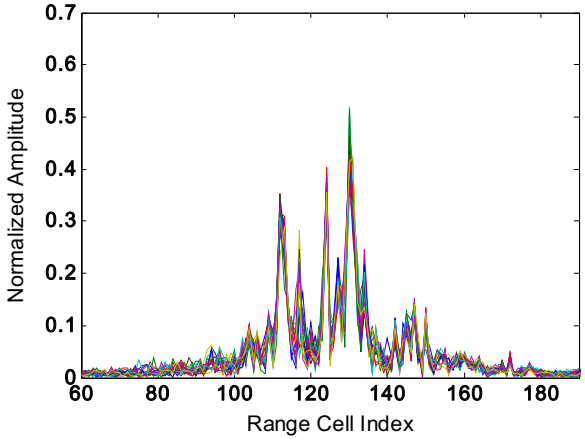

Fig. 2. Measured HRRPs with about 0.2 (degree) target-aspect variations for Yak-42 plotted with different colors. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

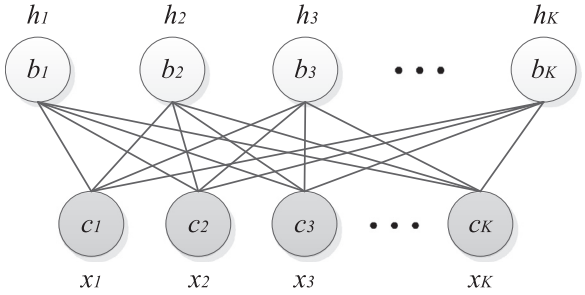

Fig. 3. The undirected graph of an RBM with K hidden and D visible units.

centers in (3), "speckle effect" arises [2,10]. A slight change in target aspect can provoke considerable variation in the phase of the scatterers, which may cause the intensity fluctuation of peaks (amplitude fluctuation property), while the location of the peaks remains unchanged (structure similarity), as shown in Fig. 2.

In practice, the performance of RATR may be heavily dependent on how to appropriately deal with or utilize those target properties [2,6,8].

2.2. Restricted Boltzmann Machine

The restricted Boltzmann machine (RBM) is a two-layer, bipartite, undirected graphical model with a set of binary hidden units $\mathbf{h}\in\{0,1\}^k$, a set of binary visible units $\mathbf{x}\in\{0,1\}^D$, and symmetric connections between these two layers represented by a weight matrix $\mathbf{W}$, see Fig. 3. For HRRP-based RATR problem, $\mathbf{x}$ can be an HRRP sample or its feature vector.

The probabilistic semantics $P(\mathbf{x},\mathbf{h})$ for an RBM is defined by its energy function $E(\mathbf{x},\mathbf{h})$ as follows:

$$P(\mathbf{x},\mathbf{h})=\frac{1}{Z}\exp\big(-E(\mathbf{x},\mathbf{h})\big) \tag{4}$$

$$E(\mathbf{x},\mathbf{h})=-\mathbf{c}^{\mathrm{T}}\mathbf{x}-\mathbf{b}^{\mathrm{T}}\mathbf{h}-\mathbf{x}^{\mathrm{T}}\mathbf{W}\mathbf{h} \tag{5}$$

where the partition function Z is given by summing over all possible pairs of visible and hidden vectors, i.e., $Z=\sum_{\mathbf{x}}\sum_{\mathbf{h}}\exp\big(-E(\mathbf{x},\mathbf{h})\big)$. $\mathbf{W}\in\mathbb{R}^{D\times K}$ is the weight matrix, $\mathbf{b}\in\mathbb{R}^K$ is the hidden bias vector, and $\mathbf{c}\in\mathbb{R}^D$ is visible bias vector.

Whereas the original RBM define the state of each visible unit to be binary, which is inconvenient for modeling real-valued data such as HRRP data. To model real-valued data, we replace the model's energy function with

$$E(\mathbf{x},\mathbf{h})=\sum_{i=1}^{D}\frac{(x_i-c_i)^2}{2\sigma_i^2}-\sum_{j=1}^{K}b_jh_j-\sum_{i=1}^{D}\sum_{j=1}^{K}\frac{x_i}{\sigma_i}h_jW_{ij} \tag{6}$$

This type of model is explored in [22,29]. The subscript i means the ith element in the vector. Here, x_i is the real-valued visible unit. Each visible unit adds a parabola (quadratic) offset to the energy function, where σ_i controls the width of the parabola.

Since there no connections between units in the same layer, visible units are conditionally independent given the hidden units (and vice versa). The conditional probabilities of individual x_i and h_j can be explicitly written as follows (see Appendix A):

$$P(x_i=x|\mathbf{h})=\mathbf{N}\left(x|c_i+\sigma_i\sum_j h_jW_{ij},\ \sigma_i^2\right) \tag{7}$$

$$P(h_i=1|\mathbf{x})=\mathrm{sigm}\left(b_j+\sum_i\frac{x_i}{\sigma_i}W_{ij}\right) \tag{8}$$

where $\mathfrak{N}\big(\cdot|\mu,\sigma^2\big)$ denotes the Gaussian probability density function with mean μ and variance σ^2. $\mathrm{sigm}(x)=\frac{1}{1+\exp(-x)}$ is the sigmoid activation function.

Given a set of N training cases $\mathbf{X}=\{\mathbf{x}_n\}_{n=1}^N$, the goal is to maximize the log-likelihood function of the set under the model's distribution:

$$\begin{aligned}\sum_{n=1}^{N}\log p(\mathbf{x}_n)&=\sum_{n=1}^{N}\log\left(\frac{1}{Z}\sum_{\mathbf{h}}e^{-E(\mathbf{x},\mathbf{h})}\right)\\&=\sum_{n=1}^{N}\left[\log\left(\sum_{\mathbf{h}}e^{-E(\mathbf{x},\mathbf{h})}\right)-\log\left(\sum_{\mathbf{x}}\sum_{\mathbf{h}}e^{-E(\mathbf{x},\mathbf{h})}\right)\right]\end{aligned} \tag{9}$$

We can optimize this by performing stochastic gradient ascent. The gradient of log-likelihood function (9) with respect to the parameter set $\Theta_{GRBM}=\{\mathbf{W},\mathbf{b},\mathbf{c},\boldsymbol{\sigma}\}$ can be computed as follows:

$$\begin{aligned}\frac{\partial\log p(\mathbf{x})}{\partial\Theta_{GRBM}}&=\frac{\partial\log\left(\sum_{\mathbf{h}}e^{-E(\mathbf{x},\mathbf{h})}\right)}{\partial\Theta_{GRBM}}-\frac{\partial\log\left(\sum_{\mathbf{x}}\sum_{\mathbf{h}}e^{-E(\mathbf{x},\mathbf{h})}\right)}{\partial\Theta_{GRBM}}\\&=-\frac{1}{\sum_{\mathbf{h}}e^{-E(\mathbf{x},\mathbf{h})}}\sum_{\mathbf{h}}\left(e^{-E(\mathbf{x},\mathbf{h})}\frac{\partial E(\mathbf{x},\mathbf{h})}{\Theta_{GRBM}}\right)\\&\quad+\frac{1}{\sum_{\mathbf{x}}\sum_{\mathbf{h}}e^{-E(\mathbf{x},\mathbf{h})}}\sum_{\mathbf{x}}\sum_{\mathbf{h}}\left(e^{-E(\mathbf{x},\mathbf{h})}\frac{\partial E(\mathbf{x},\mathbf{h})}{\Theta_{GRBM}}\right)\\&=-\mathbf{E}_{p(\mathbf{h}|\mathbf{x})}\left[\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial\Theta_{GRBM}}\right]+\mathbf{E}_{p(\mathbf{x},\mathbf{h})}\left[\frac{\partial E(\mathbf{x},\mathbf{h})}{\partial\Theta_{GRBM}}\right]\end{aligned} \tag{10}$$

Unfortunately, computing the gradient of the log-likelihood is intractable. Instead, we can use the contrastive divergence (CD) approximation through Gibbs sampling for optimizing the parameters accordingly [20,29,31,32].

A series of RBMs (GRBMs) can be then stacked to form deep networks called DBN by using the hidden activations of the previous layer as the input of the next layer.

2.3. Autoencoder

We briefly specify the autoencoder (AE) framework and its terminology here.

Encoder: The deterministic mapping that transforms an input vector $\mathbf{x}$ into hidden representation $\mathbf{h}\in\mathbb{R}^K$ is called the encoder. Its typical form is an affine mapping followed by a nonlinearity:

$$\mathbf{h}=f(\mathbf{z}_2)=\mathrm{sigm}\big(\mathbf{W}^{\mathrm{T}}\mathbf{x}+\mathbf{b}\big) \tag{11}$$

where $\mathbf{z}_1=\mathbf{x}$, $\mathbf{z}_2=\mathbf{W}^{\mathrm{T}}\mathbf{x}+\mathbf{b}$, $\mathbf{W}$ is a $D\times K$ weight matrix and $\mathbf{b}$ is an offset vector of dimensionality K.

Decoder: The resulting hidden representation $\mathbf{h}$ is then mapped back to a reconstructed D-dimensional vector $\hat{\mathbf{x}}$ in input space. Its typical form is again an affine mapping optionally followed by a squashing non-linearity, that is, either $\hat{\mathbf{x}}=\mathbf{W}\mathbf{h}+\mathbf{c}$ or

$$\hat{\mathbf{x}}=g(\mathbf{z}_3)=\mathrm{sigm}(\mathbf{W}\mathbf{h}+\mathbf{c}) \tag{12}$$

Here $\mathbf{z}_3=\mathbf{W}\mathbf{h}+\mathbf{c}$, $\mathbf{W}$ is a tied weight matrix, i.e., using the same weights for encoding the input and decoding the hidden representation.

However, without any additional constraints, traditional antoencoders can learn the identity mapping [18,21]. This phenomenon can be circumvented by using Denoising Autoencoder (DAE), which is a simple variant of the basic autoencoder described

above. It performs as well as or even better than RBM [31]. Training involves the reconstruction of a clean input from a destroyed one. Input $\mathbf{x}$ becomes corrupted input $\mathbf{y}$ by artificially adding isotropic Gaussian noise, or masking noise, or salt-and pepper noise to the clean data.

$$\mathbf{y} = \eta(\mathbf{x}) = \mathbf{x} + \mathbf{n} \tag{13}$$

A DAE is trained to denoise the input by first finding the hidden representation $\mathbf{h} = \mathrm{sigm}\left(\mathbf{W}^{\mathrm{T}}\mathbf{y} + \mathbf{b}\right)$ from which to reconstruct the original input $\hat{\mathbf{x}} = \mathrm{sigm}(\mathbf{W}\mathbf{h} + \mathbf{c})$ or $\hat{\mathbf{x}} = \mathbf{W}\mathbf{h} + \mathbf{c}$. Let $\mathbf{X} = \{\mathbf{x}_n\}_{n=1}^{N}$ represent a set of N training samples, AE (DAE) can be trained with several optimization methods to minimize the reconstruction loss:

$$L(\mathbf{X}, \Theta_{DAE}) = \arg\min_{\Theta_{DAE}} \sum_{n=1}^{N} \|\mathbf{x}_n - \hat{\mathbf{x}}_n\| \tag{14}$$

Parameter set $\Theta_{DAE} = \{\mathbf{W}, \mathbf{b}, \mathbf{c}\}$ is initialized at random and then optimized by stochastic gradient descent. According to backpropagation algorithm, the update rules for the parameters with one mini-batch are, then,

$$\nabla\mathbf{W} = \left\langle \mathbf{y}\cdot\left(\boldsymbol{\delta}^2\right)^{\mathrm{T}} + \boldsymbol{\delta}^3\cdot\mathbf{h}^{\mathrm{T}} \right\rangle \tag{15}$$

$$\nabla\mathbf{b} = \left\langle f'(\mathbf{z}_2) \odot \left(\mathbf{W}^{\mathrm{T}}\cdot\boldsymbol{\delta}^3\right) \right\rangle \tag{16}$$

$$\nabla\mathbf{c} = \left\langle g'(\mathbf{z}_3) \odot \left(g(\mathbf{z}_3) - \mathbf{x}\right) \right\rangle \tag{17}$$

where $\boldsymbol{\delta}^2 = \nabla_{\mathbf{z}_2} L = f'(\mathbf{z}_2) \odot \left(\mathbf{W}^{\mathrm{T}}\cdot\boldsymbol{\delta}^3\right)$ and $\boldsymbol{\delta}^3 = \nabla_{\mathbf{z}_3} L = g'(\mathbf{z}_3) \odot \left(g(\mathbf{z}_3) - \mathbf{x}\right)$. Here $\odot$ denotes element-wise product, $\langle\cdot\rangle$ is the average over the mini-batch of the derivative of the objective with respect to the parameters.

After finishing a DAE, we can move on to train the next layer by using the hidden layer activation of the first layer as the input of the next layer. This is usually called Stacked Denoising Autoencoders (SDAE).

In the next section, we will explain some existing issues in the practical applications and design a deep network for HRRP-based RATR.

3. Deep networks for radar HRRP target recognition

As discussed above, conventional deep networks process each HRRP sample independently, which ignore target's structure similarity and amplitude fluctuation property in HRRPs. To deal with these issues, a novel deep network referred to as Stacked Corrective Autoencoders (SCAE), is developed. The proposed model is aimed at yielding more abstract and useful hierarchical features, which can efficiently address speckle effect and outliers.

3.1. Preprocessing and the average profile

We utilize the original HRRP data as the input into deep networks directly due to their completeness and accuracy of the structural information.

Several issues should be considered when HRRP is applied to radar target recognition task. The first one is the time-shift sensitivity of HRRP. In order to decrease the computation complexity, HRRP is only a part of received radar echo extracted by a range window, in which a target signal is included. Therefore the position of the target signal in HRRP can vary with the measurement. Feature learning needs all the training samples to learn a uniform parameter model. For that reason, the absolute alignment methods [17] are good choice since their capability to process each HRRP individually. Actually, we also try some other alignment methods [2,7,13] and find that the recognition performances of different alignment methods are similar. Thus, we adopt centroid alignment [17] as the time-shift compensation technique in this paper since it is simple and easy use.

The amplitude-scale sensitivity of HRRP comes from the fact that the intensity of an HRRP is a function of radar transmitting power, target distance, radar antenna gain, radar receiver gain, radar system losses and so on. HRRPs measured by different radars or under different conditions will have different amplitude-scales. To deal with amplitude scale sensitivity, each HRRP is normalized by dividing its l_2-norm.

According to the scattering center target model[1] [2,3,13–15], the variation of the target aspect will lead to different range shifts for different scattering centers on the target, even within the aspect region where the scattering center structure remains unchanged. This is referred to as the target-aspect sensitivity. Actually, it is similar to 'coherent speckles' in SAR image [6,8,15]. In SAR imaging, the average processing is a common way to reduce the 'coherent speckles'. Obviously, the training samples in a frame for RATR using real HRRPs are more than the SAR images which can be used to do average processing. An HRRP frame here is a small target-aspect sector without scatterers' motion through range cells (MTRC) [2,3,6]. The limitation of the target-aspect change to avoid scatterers' MTRC is $\Delta\varphi \le \Delta R/L$, where ΔR denotes the length of range resolution cell and L is the cross length of the target [6,15]. For the measured ISAR data used in the paper, radar center frequency is 5520 MHz, radar bandwidth is 400 $\Delta R = 0.375\ m$. Therefore, in view of a few scatterers distributed at the edge of an aircraft-like target and with weak energy, HRRPs within no more than 3° angular sector can be taken as an HRRP frame [6].

According to the literatures [2,6,9,10,15], the definition of the average profile is:

$$\mathbf{m} \triangleq \left[\frac{1}{N}\sum_{n=1}^{N} x_{n1}, \frac{1}{N}\sum_{n=1}^{N} x_{n2}, \ldots, \frac{1}{N}\sum_{n=1}^{N} x_{nD}\right]^{\mathrm{T}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \tag{18}$$

where $\mathbf{X} = \{\mathbf{x}_n\}_{n=1}^{N}$ is an HRRP frame, with the nth HRRP sample $\mathbf{x}_n = \left[x_{n1}, x_{n2}, \ldots, x_{nD}\right]^{\mathrm{T}}$. T denotes the transpose operation. Due to the time-shift sensitivity of HRRP, to obtain the average profile, the range alignment should be performed to the set of HRRPs first; in other words, the average profile is obtained from range aligned HRRPs.

Fig. 4 is the illustration of a set of HRRPs and their corresponding average profile. It shows that the HRRPs with a large target-aspect variation differ largely, while the average profile has a smoother and concise signal form and can better represent the scattering property of the target in the given aspect-frame. Xing et al. [15] studies the average profile's physical mechanism to improve the aspect stability more detail. From the mathematic point of view, the average profile is the optimal template of the clustering center under Euclidean distance criterion [2,6]. From the perspective of signal processing, the average profile represents target's stable physical structure information in a frame [2,3,6]. An important characteristic of the average profile is that it can depress the speckle effect of HRRPs [6]. It also suppresses the impact

[1] Since targets and their components are much larger than the wavelength of microwave radar, the electromagnetism characteristics of the targets can be described by the scattering center target model. The scattering center model is a simple and efficient target approximate model (actually a first-order approximation model) in synthetic aperture radar (SAR) imaging and inverse synthetic aperture radar (ISAR) imaging.

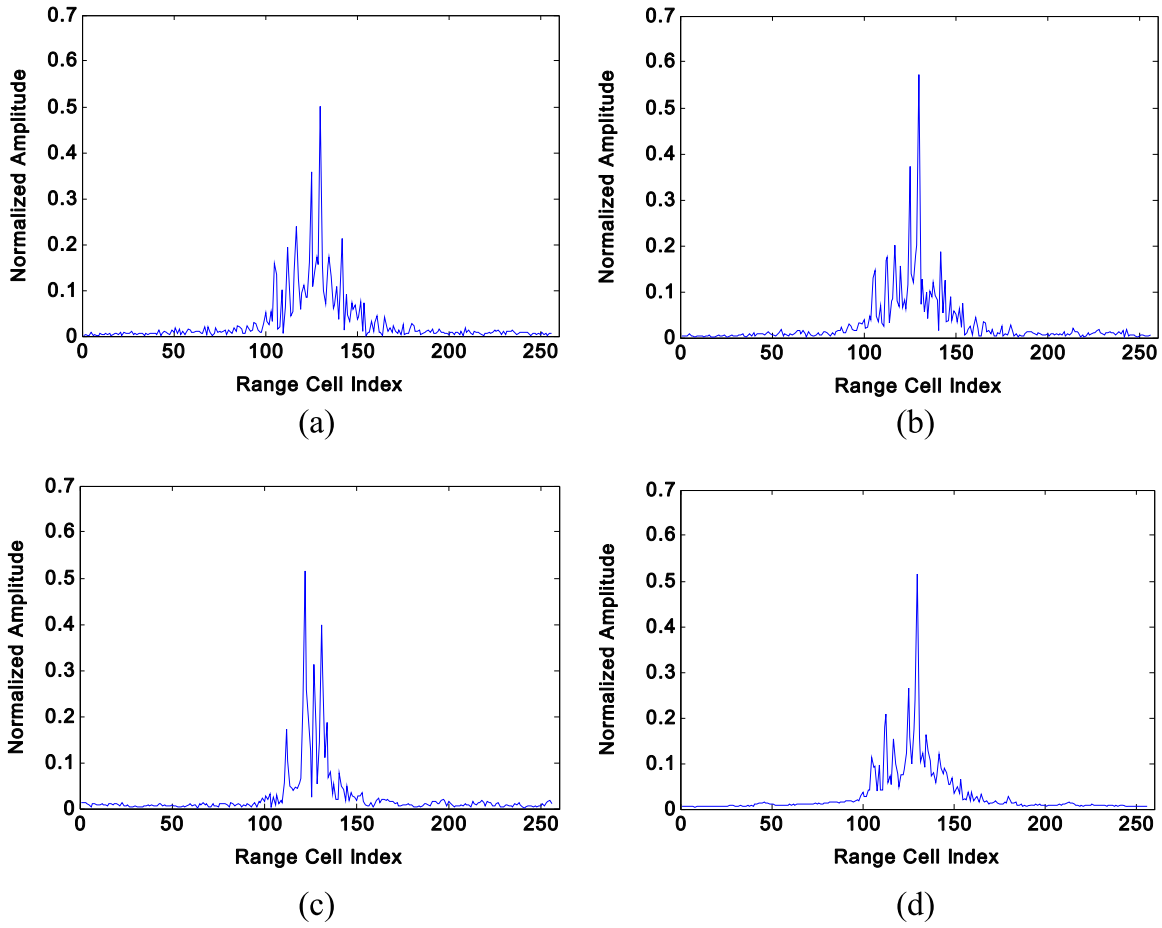

Fig. 4. The HRRP samples and the corresponding average profile of Yak-42. (a) the first HRRP. (b) the second HRRP (with a 0.01 target-aspect variation relative to the first HRRP). (c) the 350th HRRP (with a 3.5 target-aspect variation relative to the first HRRP). (d) the average profile of the top 50 HRRP samples (within a 0.5 angular sector).

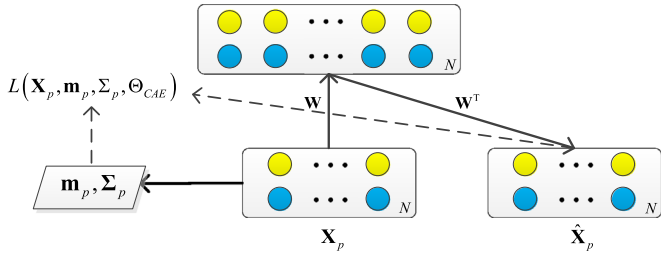

Fig. 5. Corrective Autoencoder (CAE) architecture. The number of samples in Xp is *N*. Different color represents different HRRP sample. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

of the noise spikes and the amplitude fluctuation property, which widely exists in the wideband detection and recognition [9,10].

In summary, the average profile in a small target-aspect sector has better generalization performance than the single HRRP [2,6,9,10,15]. In the following part, we will formulate a network embedded with average processing for HRRP-based RATR.

3.2. Stacked CAEs for Radar HRRP Target Recognition

3.2.1. Corrective Autoencoder (CAE)

In HRRP-based RATR field, a good feature is one that can be obtained robustly from a corrupted observation and that will be useful for mapping to the desired signal. Therefore, we propose a novel framework called Corrective Autoencoder (CAE) to learn such features. CAE is a training scheme that learns a nonlinear space mapping a single HRRP sample close to its desired signal, which we named correction term. We utilize the average profile as the correction term and the single HRRP as the corrupted observation in this network. Since the covariance matrix in an HRRP frame sufficiently considers the diversity and correlation among range cells, it is embedded in CAE for establishing a more effective loss function. CAE incorporates the HRRP frame and average processing into one autoencoder. Such architecture is beneficial for preserving the target's structure information, which helps us to learn stable and useful features for RATR. The architecture of CAE is shown in Fig. 5.

Let the training dataset be $\mathbf{X} = \{\mathbf{X}_p\}_{p=1}^{P}$, where $\mathbf{X}_p = \{\mathbf{x}_{p,n}\}_{n=1}^{N}$ denotes the examples in the pth HRRP frame. We minimize the loss function under the Mahalanobis distance criterion by a sparsity-inducing term so as to combine the virtues of "sparsity" and CAE:

$$L(\mathbf{X}, \Theta_{CAE}) = \frac{1}{T}\sum_{p=1}^{P}\sum_{n=1}^{N}\frac{1}{2}\left(\mathbf{m}_p - \hat{\mathbf{x}}_{p,n}\right)^{\mathrm{T}}\boldsymbol{\Sigma}_p^{-1}\left(\mathbf{m}_p - \hat{\mathbf{x}}_{p,n}\right) + \alpha KL\left(\rho\|\hat{\boldsymbol{\rho}}\right) + \frac{\beta}{2}\|\mathbf{W}\|_{\mathrm{F}}^2 \tag{19}$$

where

$$KL\left(\rho\|\hat{\boldsymbol{\rho}}\right) = \sum_{j=1}^{|\hat{\rho}|}\rho\log\frac{\rho}{\hat{\rho}_j} + (1-\rho)\log\frac{1-\rho}{1-\hat{\rho}_j}, \qquad \hat{\boldsymbol{\rho}} \triangleq \frac{1}{T}\sum_{p=1}^{P}\sum_{n=1}^{N}\mathbf{h}_{p,n} \tag{20}$$

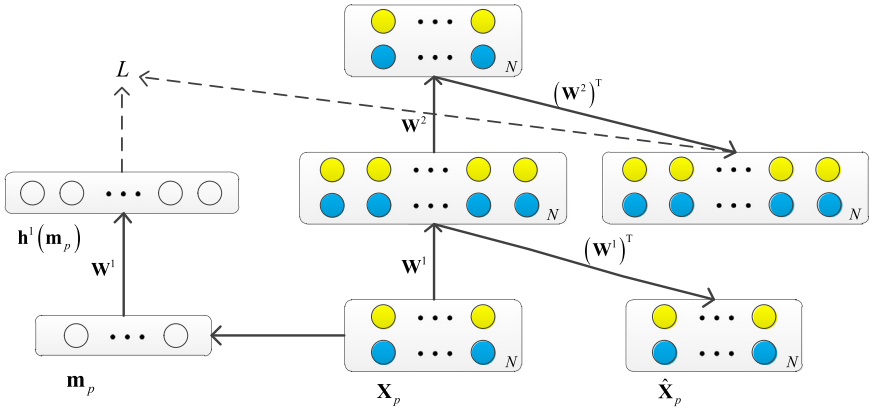

Fig. 6. Stacking Corrective Autoencoders. After training a first layer CAE (see Fig. 5) its learnt encoding weights $\mathbf{W}^1$ is used on the single HRRP and the average profile. The resulting features are used to train a second layer CAE to learn a second layer encoding weights $\mathbf{W}^2$. From there, the procedure can be repeated.

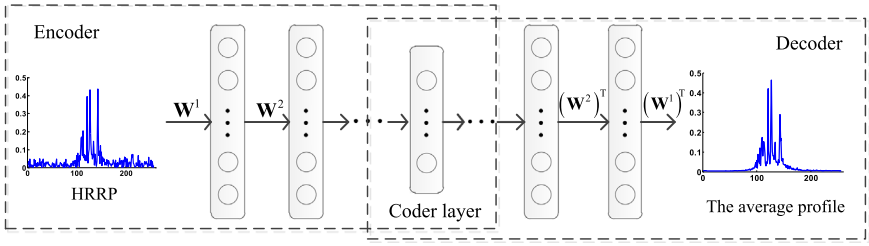

Fig. 7. Stacked Corrective Autoencoders (SCAE) architecture for HRRP-based RATR. Here Stacking CAEs are "unrolled" for visualization.

$$\hat{\mathbf{x}}_{p,n} = g(\mathbf{z}_3) = \begin{cases} \mathbf{W}\mathbf{h}_{p,n} + \mathbf{c}, & \mathbf{x}_{p,n} \in \mathbb{R}^D \\ \mathrm{sigm}\left(\mathbf{W}\mathbf{h}_{p,n} + \mathbf{c}\right), & \mathbf{x}_{p,n} \in \{0,1\}^D \end{cases} \mathbf{h}_{p,n} = f(\mathbf{z}_2)$$
$$= \mathrm{sigm}\left(\mathbf{W}^{\mathrm{T}}\mathbf{x}_{p,n} + \mathbf{b}\right) \tag{21}$$

with $\Theta_{CAE} = \{\mathbf{W}, \mathbf{b}, \mathbf{c}\}$, $\mathbf{z}_2 = \mathbf{W}^{\mathrm{T}}\mathbf{x}_{p,n} + \mathbf{b}$, $\mathbf{z}_3 = \mathbf{W}\mathbf{h}_{p,n} + \mathbf{c}$ and $T = P{\cdot}N$. $\mathbf{m}_p$ denotes the average profile in the pth HRRP frame, which is the correction term in $\mathbf{X}_p$ that enforces the reconstruction samples similar to the average profile.

$$\mathbf{m}_p = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_{p,n} \tag{22}$$

According to the literatures [4,13,14,16], each HRRP frame approximately follows independent Gaussian distribution. Since the sufficient statistic of Gaussian distribution is the mean and covariance, we further combine the covariance matrix of the each HRRP frame into the loss function (19) for completely describing the distribution structure of HRRP. The covariance matrix of the pth HRRP frame is

$$\mathbf{\Sigma}_p = \frac{1}{N}\sum_{n=1}^{N} \left(\mathbf{x}_{p,n} - \mathbf{m}_p\right)\left(\mathbf{x}_{p,n} - \mathbf{m}_p\right)^{\mathrm{T}} \tag{23}$$

We also force the features to be "sparse", in that only a tiny fraction of the hidden units should be active in relation to a given stimulus. The KL-divergence term will encourage each of hidden units to have a mean activation close to some small constant ρ. $\|\mathbf{W}\|_{\mathrm{F}}^2$ is a weight decay term which penalizes large values of the parameters.

Parameter set Θ_{CAE} is initialized at random and then optimized by stochastic gradient descent. According to backpropagation algorithm [31,32], the update rules for the parameters with one mini-batch are

$$\nabla\mathbf{W} = \left\langle \mathbf{x}_{p,n}{\cdot}\left(\boldsymbol{\delta}^2\right)^{\mathrm{T}} + \boldsymbol{\delta}^3{\cdot}\left(\mathbf{h}_{p,n}\right)^{\mathrm{T}} \right\rangle + \beta\mathbf{W} \tag{24}$$

$$\nabla\mathbf{b} = \left\langle f'(\mathbf{z}_2) \odot \left(\mathbf{W}^{\mathrm{T}}{\cdot}\boldsymbol{\delta}^3 + \alpha{\cdot}\mathbf{t}\right) \right\rangle \tag{25}$$

$$\nabla\mathbf{c} = \left\langle \mathbf{\Sigma}_p^{-1}\left(g(\mathbf{z}_3) - \mathbf{m}_p\right) \odot g'(\mathbf{z}_3) \right\rangle \tag{26}$$

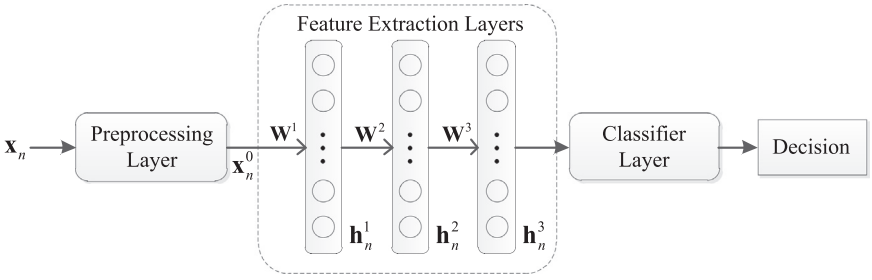

Fig. 8. Typical radar HRRP target recognition procedure with deep networks.

where

$$\boldsymbol{\delta}^2 = \nabla_{\mathbf{z}_2} L = f'(\mathbf{z}_2) \odot \left(\mathbf{W}^{\mathrm{T}} \cdot \boldsymbol{\delta}^3 + \alpha \cdot \mathbf{t}\right) \tag{27}$$

$$\boldsymbol{\delta}^3 = \nabla_{\mathbf{z}_3} L = \boldsymbol{\Sigma}_p^{-1}\left(g(\mathbf{z}_3) - \mathbf{m}_p\right) \odot g'(\mathbf{z}_3) \tag{28}$$

and the jth element of vector $\mathbf{t}$ is $t_j = -\frac{\rho}{\hat{\rho}_j} + \frac{1-\rho}{1-\hat{\rho}_j}$, $\odot$ denotes element-wise product, $\langle\cdot\rangle$ is the average over the mini-batch of the derivative of the objective with respect to the parameters.

3.2.2. Stacked Corrective Autoencoders (SCAE)

CAE is a single-layer learning module. After finishing a CAE, we extend the CAE architecture to L hidden layers, where the weights in layer l are trained by keeping all the weights in the lower layers constant and taking as data the activities of the hidden units at layer $l-1$. The input of our deep networks is a D-length HRRP sample $\mathbf{x}_{p,n}$, for $n = 1, \ldots, N$, under an arbitrary target-sensor orientation, and the output is its top-level features $\mathbf{h}_{p,n}^{L}$.

In the first layer, $\mathbf{x}_{p,n}$ is transformed to a K_1-length hidden units $\mathbf{h}_{p,n}^{1}$ through a weight matrix $\mathbf{W}^{1} \in \mathbb{R}^{D \times K_1}$. Each $\mathbf{h}_{p,n}^{1}$ can be computed as

$$\mathbf{h}_{p,n}^{1} = \mathrm{sigm}\left(\left(\mathbf{W}^{1}\right)^{\mathrm{T}} \mathbf{x}_{p,n} + \mathbf{b}^{1}\right) \tag{29}$$

Since the HRRP sample $\mathbf{x}_{p,n} \in \mathbb{R}^{D}$, the reconstruction sample $\hat{\mathbf{x}}_{p,n}$ can be

$$\hat{\mathbf{x}}_{p,n} = \mathbf{W}^{1} \cdot \mathbf{h}_{p,n}^{1} + \mathbf{c}^{1} \tag{30}$$

After training of the first CAE, we use $\mathbf{h}_{p,n}^{1}$ and $\mathbf{h}^{1}(\mathbf{m}_p)$ as the input data and the correction term for the second CAE, respectively. Each $\mathbf{h}_{p,n}^{2}$ can be computed as

$$\mathbf{h}_{p,n}^{2} = \mathrm{sigm}\left(\left(\mathbf{W}^{2}\right)^{\mathrm{T}} \mathbf{h}_{p,n}^{1} + \mathbf{b}^{2}\right) \tag{31}$$

The reconstruction is an affine mapping followed by a squashing non-linearity when the input is not real-valued, i.e., $\hat{\mathbf{h}}_{p,n}^{1} = \mathrm{sigm}\left(\mathbf{W}^{2}\mathbf{h}_{p,n}^{2} + \mathbf{c}^{2}\right)$. Here $\mathbf{h}^{1}(\mathbf{m}_p)(p = 1, \ldots, P)$ is the feature vector of the average profile $\mathbf{m}_p$.

$$\mathbf{h}^{1}(\mathbf{m}_p) = \mathrm{sigm}\left(\left(\mathbf{W}^{1}\right)^{\mathrm{T}} \mathbf{m}_p + \mathbf{b}^{1}\right) \tag{32}$$

We point out that it is more reasonable in that, since the hierarchical feature $\mathbf{h}^{1}(\mathbf{m}_p)$ extracted from the average profile retains the stable structure of the average profile, while the average features $\frac{1}{N}\sum_{n=1}^{N}\mathbf{h}_{p,n}^{1}$ does not have this property [3]. The complete architecture for learning and stacking several layers of CAEs, i.e., SCAE, is shown in Fig. 6. To illustrate this further, the deep networks for HRRP-based RATR can be "unrolled" for visualization as shown in Fig. 7.

SCAE is a training scheme based on autoencoder with the purpose of finding a robust nonlinear space that maps a single HRRP sample close to its corresponding average profile. Compared with the conventional deep networks, there are several differences between the architectures. First, the whole HRRP frame is inputted to SCAE at a time (see Figs. 5 and 6), whilst DBN and SDAE proceed using each example independently, which ignore the target's structure similarity in the same HRRP frame.

Next, in SDAE, the loss function is set up by minimizing the square error between the clear HRRP data and the artificially corrupted data. The corrupted data is usually obtained by adding the noise with certain distributions, such as isotropic Gaussian noise, masking noise, salt-and pepper noise, to the clean data [21]. However, due to the uncertainties of the target-aspect sensitivity, amplitude fluctuation, noise spikes in HRRPs, the prior distribution of these corruptions is unknown, thus it is difficult to generate the observations by adding any analytical noise on the given "clean" data. SCAE, instead, directly utilizes the average profile as the correction term and the single HRRP as the corrupted observation. The object of SCAE is to minimize the Mahalanobis distance between the single HRRP and its corrective tem, which is a more suitable criterion for HRRP-based RATR [13].

Moreover, both $\mathbf{h}_{p,n}^{l}$ and $\mathbf{h}^{l}(\mathbf{m}_p)$ are transferred to the next layer as the training data and correction term respectively in SCAE. This is also different from SDAE, where only $\mathbf{x}_{p,n}$ is passed on to the networks to produce the clean training data for the next layer while the noisy data is discarded, and the noisy data for the next layer is constructed by randomly corrupting the generated clean training data. We point out that SCAE is more natural in that, since $\mathbf{h}_{p,n}^{l}$ lies in a different space from $\mathbf{x}_{p,n}$, the meaning of applying $\eta(\cdot)$ (see (13)) to $\mathbf{h}_{p,n}^{l}$ is not clear.

Table 1
Parameters of planes and radar in the ISAR experiment.

Radar parameters	**Center frequency**	5520 MHz	
	bandwidth	400 MHz	
Planes	**Length** (m)	**Width** (m)	**Height** (m)
Yak-42	36.38	34.88	9.83
Cessna Citation S/II	14.40	15.90	4.57
An-26	23.80	29.20	9.83

4. Parameters setting and experimental results

In this section, we first study the methodology and introduce the measured HRRP data used in our experiments. Then, the influence of model parameters will be discussed to help to realize the selection of model parameters. Finally, several detailed experiments to analyze the recognition and generalization performance of algorithms are presented and studied, respectively.

4.1. Methodology

Typical radar HRRP target recognition procedure with deep networks is illustrated in Fig. 8. In the following experiments, we use this framework for all the deep networks, including DBN, SDAE and SCAE.

In the data preprocessing layer, an HRRP training sample $\mathbf{x}_n$ is transformed to $\mathbf{x}_n^0$ through the sensitivity elimination techniques mentioned above.

The feature extraction layers: The depth of a deep network, i.e., the number of the feature extraction layers in our paper, is the length of the longest path from an input node of the network to an output node of the network. The crucial property of a deep network is that its number of paths, i.e., ways to reuse different parts, can grow exponentially with its depth. [18] shows that a deep representation can be exponentially more efficient than one that is insufficiently deep. However, too many layers also make the deep architecture hard to train effectively. They bring more parameters to learn, and make the representations in the higher layers excessive abstract, which can just overfit the training set whether the lower layers compute useful features or not. Through the analysis of experiments, we find that the number of layers $L = 3$ is a good selection for our HRRP data. Therefore, we construct three fully connected layers, which encode $\mathbf{x}_n^0$ into hierarchical features $\mathbf{h}_n^3$.

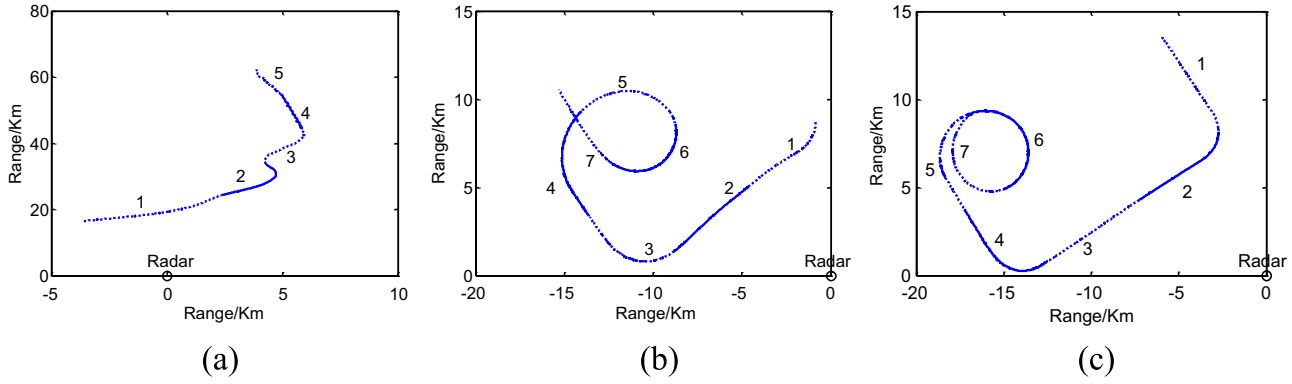

Fig. 9. Projections of target trajectories onto the ground plane. (a) Yak-42. (b) Cessna Citation S/II. (c) An-26.

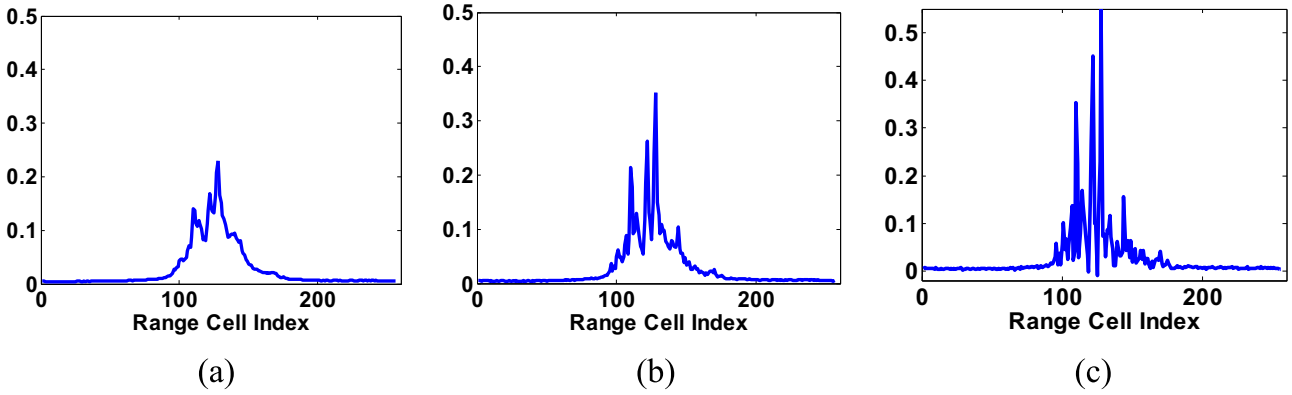

Fig. 10. The reconstruction HRRP sample via using different weights initialization scaling coefficient a. (a) $a = 0.5$. (b) $a = 1$. (c) $a = 2$.

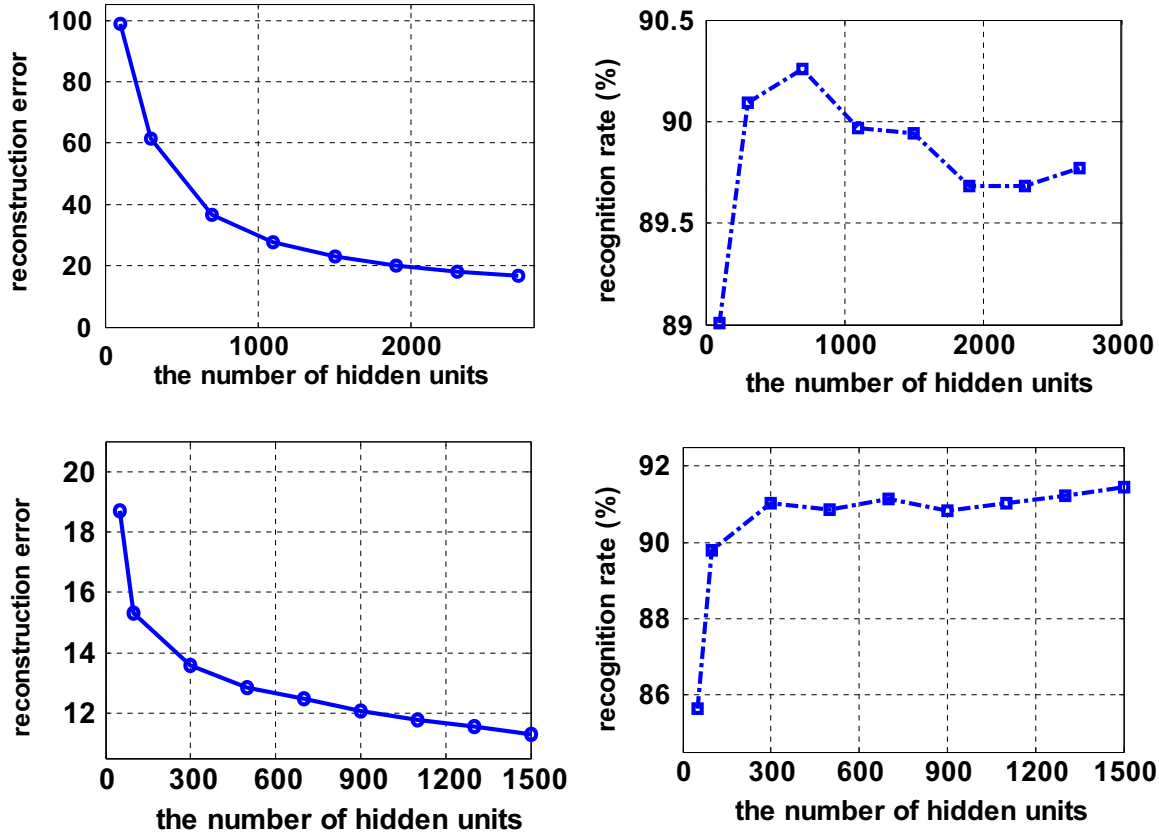

Fig. 11. Variation of the reconstruction error and recognition rate versus the number of hidden units in layer one (top row) and layer two (bottom row).

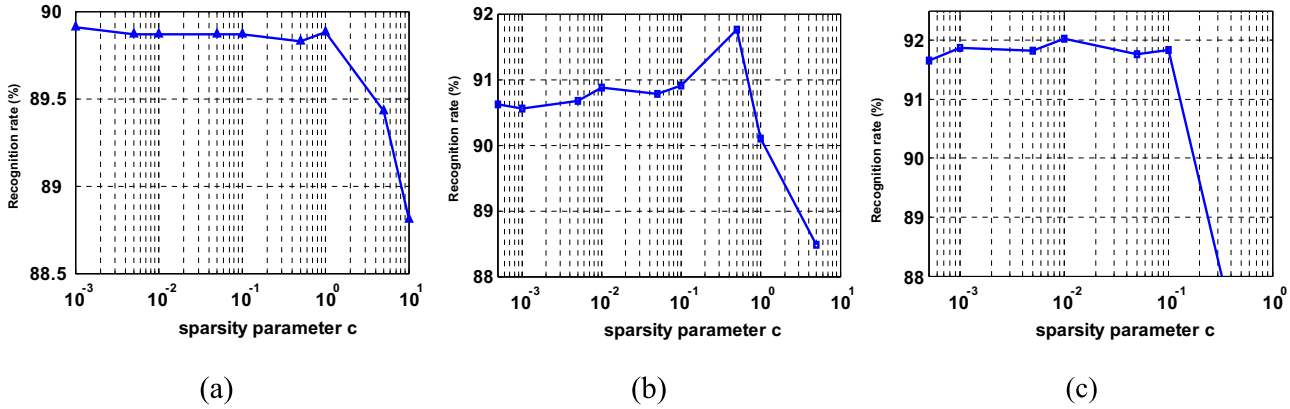

Fig. 12. Variation of the recognition accuracies versus the sparsity parameter *c*. (a) layer one. (b) layer two. (c) layer three.

Table 2
Classification performance of the proposed model (denoted SCAE, for Stacked Corrective Autoencoders) with several traditional methods. The size of deep architecture is 256-1500-500-50.

Input	Single HRRP	Average profile
LDA	81.3%	73%
K-SVD	74.7%	70.83%
PCA	83.81%	83.66%
MCC	62.42%	63.76%
AGC	85.63%	77.09%
LSVM	86.7%	88.28%
DBN Layer-1	90.04%	89.96%
DBN Layer-2	90.29%	90.64%
DBN Layer-3	89.29%	89.65%
SDAE Layer-1	88.31%	89.68%
SDAE Layer-2	90.29%	91.10%
SDAE Layer-3	90.42%	91.20%
SCAE Layer-1	89.93%	
SCAE Layer-2	91.56%	
SCAE Layer-3	***92.03%***	

For DBN, the first feature extraction layer of DBN is a GRBM for modeling real-valued HRRP data. The second and third layers are RBMs for modeling the binary hidden units in the lower layers. We use the CD-1 approximation [20] through Gibbs sampling for optimizing all the parameters.

For SDAE, only the clean data is provided while the noisy version of it is generated during training by adding real random Gaussian to the clean data. Note that there is a slight difference in the first DAE. We denote $\tilde{\mathbf{x}}_n$ as the complex returned echo vector. According to the definition of HRRP in the earlier literatures, an HRRP $\mathbf{x}_n$ is a real vector, which is the amplitude of complex returned echo vector in baseband, i.e., $\mathbf{x}_n = |\tilde{\mathbf{x}}_n|$. Therefore in the first layer, $\mathbf{x}_n$ is the clean sample, and the corresponding artificially corrupted sample is $|\eta(\tilde{\mathbf{x}}_n)| = |\tilde{\mathbf{x}}_n + \mathbf{n}|$ ($\mathbf{n}$ here is a random complex Gaussian white noise). After training of the first DAE, we use the resulting features as the clean input for the second DAE, and so on.

Finally, once a stack of networks has been built, its top-level output can be used as the input to a classifier, for example an linear support vector machine (LSVM) [18,21,26] or a (multi-class) logistic regression [20,27–29]. LSVM is an efficient machine learning algorithm aimed at minimizing structural risk for good generalization performance. Its linear structure ensures a better ability to deal with large-scale data. In the follow-up experiments, the top-level features $\mathbf{h}_n^3$ are taken as the input pattern to train an LSVM classifier in order to reflect the linear separability of the extracted features.

4.2. Measured HRRP data

To demonstrate the characteristic of the models, we consider measured radar HRRP data from three real airplanes, which are extensively used in [1–3,5–7,13–17]. Yak-42 is a large and medium-sized jet aircraft, Cessna Citation S/II is a small-sized jet aircraft and An-26 is a medium-sized propeller aircraft. The specific parameters of the airplane targets and measurement radar are shown in Table 1 and the projections of target trajectories onto the ground plane are shown in Fig. 9, from which the aspect angle of an airplane can be estimated according to its relative position to radar.

According to the literatures [1–3,13], we choose the training and test datasets with two preconditions: (a) The training dataset cover almost all of the target-aspectangles of the test dataset; (b) The elevation angles of the test dataset are different from those of the training dataset. Therefore, data from the 2th and the 5th segments of Yak-42, the 6th and the 7th segments of Cessna Citation S/II, the 5th and the 6th segments of An-26 are taken as the training samples, and other data segments are taken as test samples. There are 2800 HRRP frames for training, each of which has 50 adjacent HRRP samples. The average profile can be calculated by the mean of the aligned HRRPs in each frame. There are 700 average profiles for Yak-42, 1050 for Cessna, and 1050 for An-26. The target-aspect change between the two neighboring HRRPs is

Table 3
The confusion matrices of Three deep networks. The Size of Deep Architecture is 256-1500-500-50. The input of SDAE and DBN is the single HRRP.

Methods	DBN			SDAE			SCAE		
	Yak-42	Cessna	An-26	Yak-42	Cessna	An-26	Yak-42	Cessna	An-26
Yak-42	97.08	0.58	2.33	96.90	0.50	2.60	97.33	1.33	1.33
Cessna	1.60	83.75	14.65	0.75	86.36	12.89	1.35	91.75	6.90
An-26	8.75	4.21	87.04	7.75	4.25	88.00	9.25	3.75	87.00
Av. R	89.29			90.42			**92.03**		

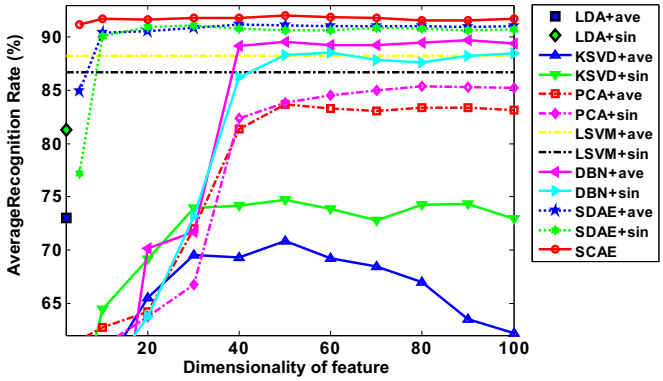

Fig. 13. Variation of the recognition performance with dimensionality of feature, via four shallow models: LDA, K-SVD, PCA, LSVM and three deep models: DBN, SDAE and SCAE. We use a 256-1500-500-50 network for all the deep models.

about 0.01°. The number of test samples is 5200. Each sample is a 256-dimensional vector, i.e., $D = 256$. The number of target classes is C=3.

4.3. The influence of model parameters

Although the deep architectures constructed in Section 3 may appear relatively complex, the number of parameters that need tuning is not particularly large to avoid overfitting, and they are set in a "standard" way [33,34]. In the following experiments, all the deep networks have $L = 3$ fully connected code layers, which encode $\mathbf{x}_n$ into the hierarchical features $\mathbf{h}_n^3$. In this subsection, we analyze the influence of parameters at each of the levels of the hierarchy on the performance.

Weights initialization scaling coefficient: Weights need to be initialized carefully to break the symmetry between hidden units of the same layer. Hinton and Bengio [33,34] and the Deep Learning Tutorials recommend to use a combination of the fan-in and fan-out, e.g., sample a Uniform $(-r, r)$ with $r = a\cdot\sqrt{6/(\text{fan-in} + \text{fan-out})}$. In Fig. 10, we show that $a = 1$ is a better choice for HRRP-based RATR. When the value of a is much larger, CAE may reconstruct more detail noise; when a is smaller, CAE will lose more signal information.

The number of hidden units: The reconstruction performance can be improved with the increase of the number of hidden units, whereas the larger size of hidden units will lead to higher computational complexity, and cannot always keep the recognition accuracy improving, as shown in the right column of Fig. 11. Here for clear and simple illustration, we do not show the results of the third layer, instead, the number of layer-3 hidden units is fixed (e.g., $K_3 = 50$) in all our experiments.

Besides the accuracy, we would like to also check the influence of the number of hidden units on the reconstruction error, which can be calculated at layer l by

$$err_l = \frac{1}{N\cdot P}\sum_{p=1}^{P}\sum_{n=1}^{N}\|\mathbf{h}_{p,n}^{l-1} - \hat{\mathbf{h}}_{p,n}^{l-1}\|_2^2, \qquad 1 < l \le L \tag{33}$$

$$\mathbf{h}_{p,n}^{l-1} = \text{sigm}\left(\left(\mathbf{W}^{l-1}\right)^{\mathrm{T}}\mathbf{h}_{p,n}^{l-2} + \mathbf{b}^{l-1}\right), \quad \mathbf{h}_{p,n}^{0} \triangleq \mathbf{x}_{p,n} \tag{34}$$

$$\hat{\mathbf{h}}_{p,n}^{l-1} = \text{sigm}\left(\mathbf{W}^{l}\mathbf{h}_{p,n}^{l} + \mathbf{c}^{l}\right), \tag{35}$$

where $\mathbf{h}_{p,n}^{l-1}$ is the nth feature sample of layer (l-1) in the pth HRRP frame, $\hat{\mathbf{h}}_{p,n}^{l-1}$ is the corresponding reconstruction feature vector. At layer $l = 1$ (Row 1 in Fig. 11), we need to consider the trade-off between the reconstruction and recognition performance, thus $K_1 = 1500$ is a good choice. At layer $l = 2$ (Row 2 in Fig. 11), both the reconstruction and recognition performance are better when $K_2 \ge 500$. So we choose $K_2 = 500$ for a lower computational complexity.

Sparisity: As discussed in [26,32–35], a sparsity-inducing penalty is a way to regularize, which may be advantageous for the deep networks.

Fig. 12 shows the variation of average recognition accuracies versus the sparsity parameter c. At layer $l = 1$ (Fig. 12(a)), it indicates that we do not need too large sparsity penalty because the weights elements are fundamental (primitive). While with increasing l, the weight elements become more specialized and therefore sparsity is useful (Fig. 12(b and c)). Note that, in our experiments, we fix a small value $K_3 = 50$ as the number of hidden units at layer $l = 3$ for reducing the dimensionality of data. Thus, the recognition performance has somewhat improvement when using sparsity penalty ($c = 0.01$ in Fig. 12(c)).

4.4. Recognition performance

According to the conclusion above, we set the number of hidden units in each layer are $K_1 = 1500$, $K_2 = 500$ and $K_3 = 50$, respectively.

In Table 2, we show the superior recognition performance of SCAE, where SCAE is compared with four shallow models, including Linear discriminant analysis (LDA), K-SVD, Principal component analysis (PCA), and LSVM, and two deep networks, i.e., DBN and SDAE. The last column of Table 2 is completed by using the average profiles obtained by (18) as the inputs to train the models. Note that both the single HRRP data and the average profile are taken into consideration in SCAE. The feature dimension of LDA is C – 1 according to the literature [30]. The feature dimension of KSVD is set to 1500 with no more than 50 nonzero elements, which is as same as the parameter setting in the first layer of deep networks. We apply other models to obtain compact features of 50 dimensions, and then using these features as the input to the classifier. LSVM here only serves as a simple baseline, thus it does not employ any feature extraction. Some statistical recognition methods for HRRP, e.g., Maximum correlation coefficient (MCC) [6], Adaptive Gaussian classifier (AGC) [13], are also considered here to demonstrate our method's validity.

Table 4
Classification Performance of The Proposed Model (denoted SCAE, for stacked corrctive autoencoders) with Several Traditional Methods in the Case of Half Training Dataset. The Size of Deep Architecture is 256-1300-200-50.

Input	Single HRRP	Average profile
LDA	70.93%	66.12%
K-SVD	60.92%	51.44%
PCA	72.6%	73.02%
MCC	56.13%	57.03%
AGC	74.37%	63.86%
LSVM	80.1%	79.01%
DBN Layer-1	82.47%	81.47%
DBN Layer-2	83.07%	82.63%
DBN Layer-3	82.23%	83.54%
SDAE Layer-1	81.37%	81.03%
SDAE Layer-2	83.33%	83.07%
SDAE Layer-3	84.26%	83.17%
SCAE Layer-1	82%	
SCAE Layer-2	85.17%	
SCAE Layer-3	***85.64%***	

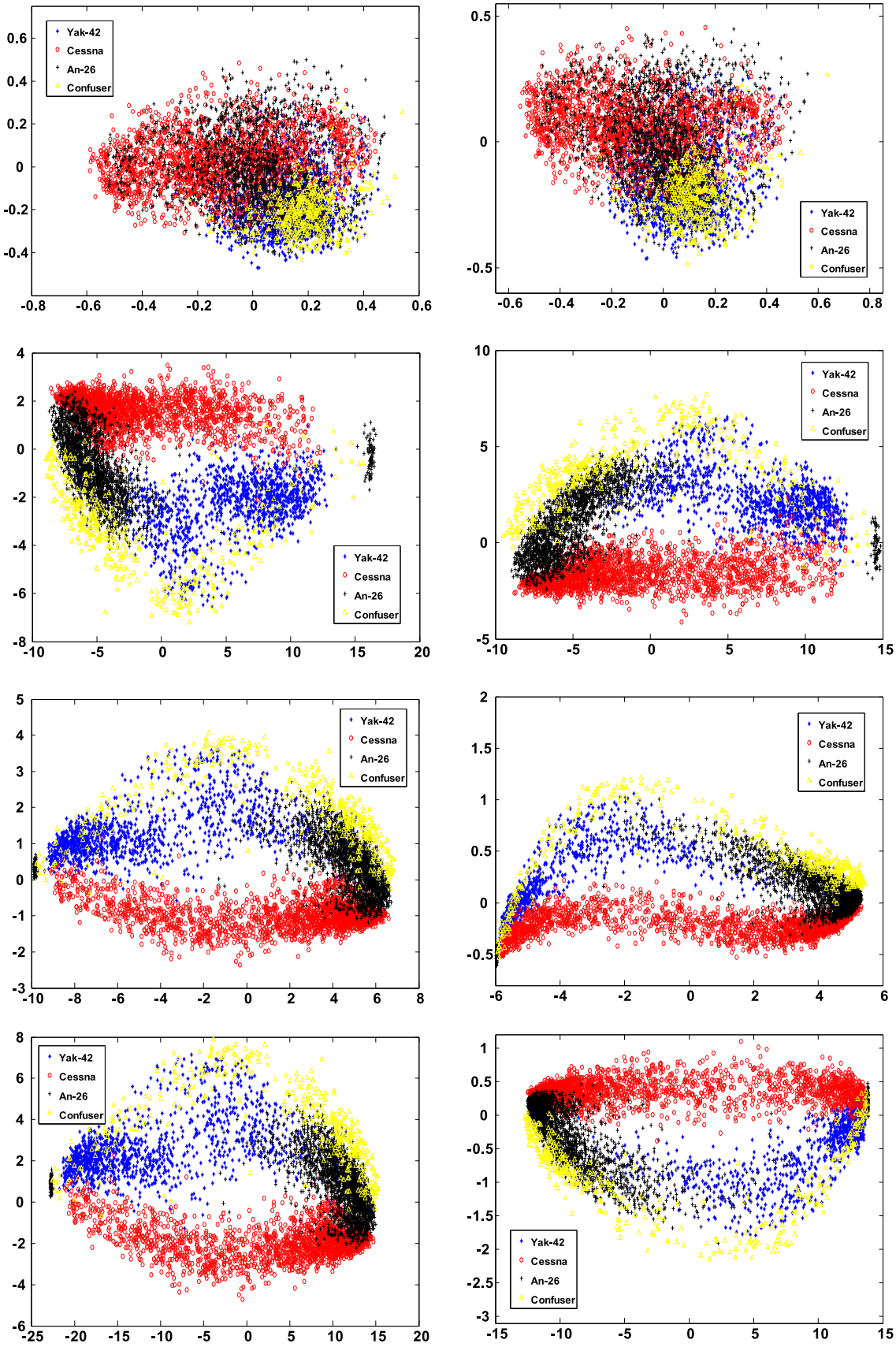

Fig. 14. Visualizations of the test HRRP samples and their corresponding features in different layers, via using two-dimensional PCA. **Column** 1: Complete training dataset. **Column** 2: Incomplete training dataset. **Row** 1: The original HRRP. **Row** 2: The features in the first-layer. **Row** 3: The features in the second-layer. **Row** 4: The features in the third-layer.

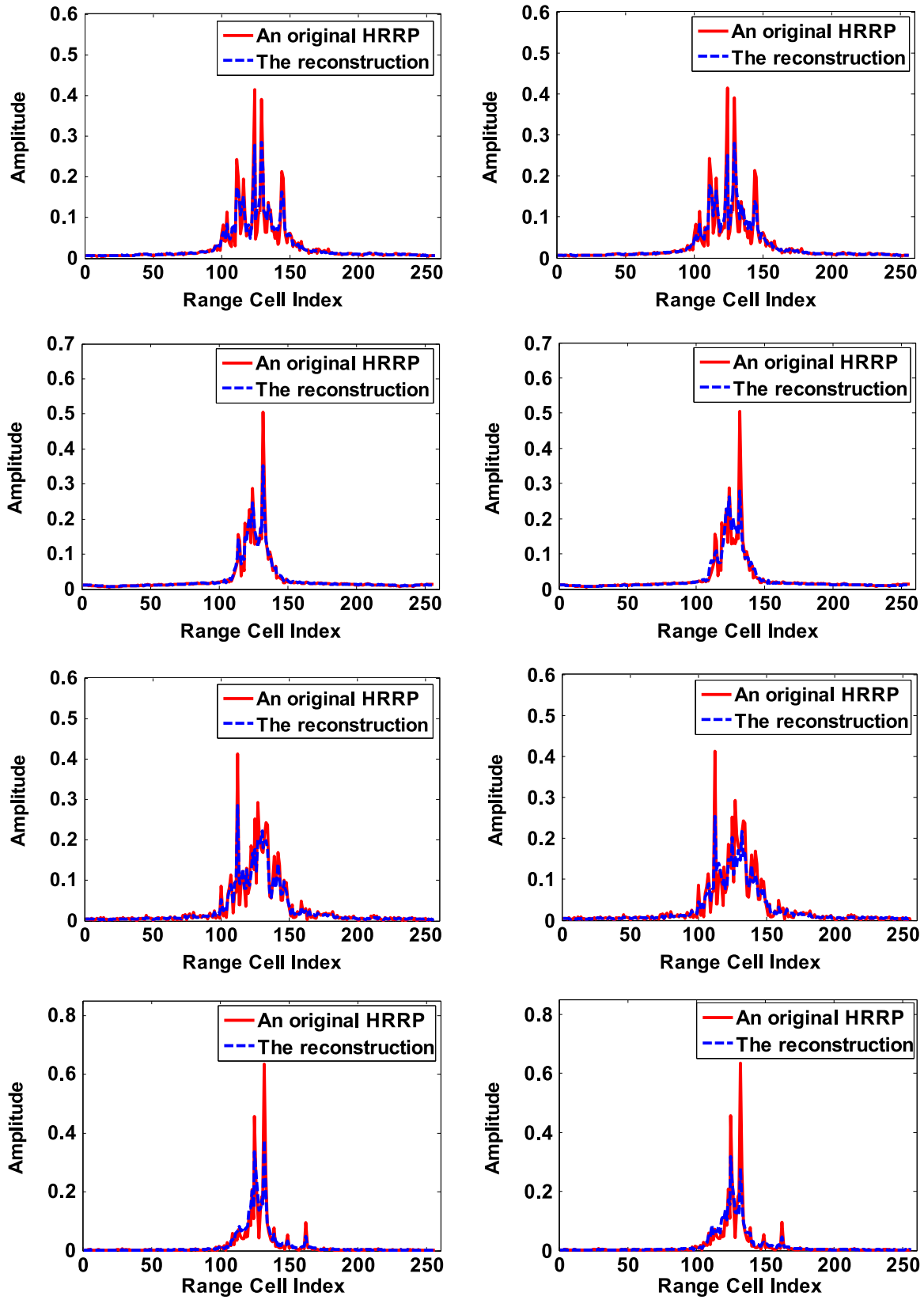

Fig. 15. The reconstruction performance of SCAE for four HRRP samples. **Column** 1: Complete training dataset. **Column** 2: Incomplete training dataset. **Row** 1: A training sample in the incomplete training dataset. **Row** 2: A training sample out of the incomplete training dataset. **Row** 3: A test sample. **Row** 4: A sample in the confuser target dataset.

Several observations can be made from Table 2. First, the recognition accuracy of PCA using single HRRP as the input is 83.81%, which is the best result in the shallow non-statistical models but still less than the recognition rate of the deep networks about 6–8%. Therefore, compared with the shallow architecture, the nonlinear hierarchical architecture of deep networks is more

advantageous to learn the effective distinguishing features, and results in significantly performance improvement for RATR. Next, among those deep networks, SCAE has the best recognition performance, and the higher level features in SCAE are more effective relative to the first layer alone. That means SCAE can map the original input to a new and highly abstract feature space layer by layer, and facilitate the RATR task as hierarchy is ascended. For statistical recognition methods, MCC calculates the correlation coefficient between the test sample and each template [6]. Such shallow model is simple but has very limited recognition performance. AGC is a more complex statistical model with more parameters, which assumes the sample obeys an independent Gaussian distribution. However, when the training data size is small or the test sample does not obey the learned training data distribution, i.e., there are differences between the test sample and training data, the recognition performance will degrade.

To look insight into the performance of the methods on different targets, we also list the confusion matrix for DBN, SDAE and SCAE (Table 3), since they are leading ones. The recognition performance of SCAE raising mainly come from the improvement in Cessna, while the accuracy of An-26 just has a slight drop. Through the analysis of the measured data, we find that the waveform of Cessna is not smooth and usually has larger fluctuation. SCAE, which makes full use of the connections between the rough single HRRP and smooth average profile coherently in a unified framework, may depress such fluctuation and increase the classification performance. Moreover, in SCAE, the feature dimension of the classifier's input is 50, while the best performance achieved by LSVM is 88.28% with the original number of features (256), which is inferior to SCAE about 3.8%. Therefore, SCAE is also a good dimension reduction method, which will be analyzed in detail in the following experiment.

In order to investigate the inference of the feature dimensionality, we also plot the recognition performances of each method varying with the final dimensionality of extracted features in Fig. 13. From Fig. 13, we can see that deep networks works more stable and better than other traditional models, especially when the dimensionality is small. We attribute the reason to the multilayer structure of the deep models, which can capture more discriminative features consistent across the training data. And SCAE gives the leading performance, since SCAE makes full use of the structural stability of the average profiles and depresses the harmful fluctuations caused by the target structure. Therefore, deep models can utilize only a few of features to characterize the differences between categories, and the learned features are more effective for recognition and facilitate the visualization, storage for RATR. Note that, due to the unstable recovery of sparse overcomplete representation [35], there is a decrease in the KSVD's recognition accuracy when the number of nonzero feature elements is larger than 50.

In summary, SCAE is a more suitable deep architecture to obtain better recognition performance for HRRP-based RATR.

4.5. Generalization performance

In aforementioned experiments, the training data used to train the model covers all possible target-aspect. However, in the noncooperative circumstance, e.g., at the battle time, the training dataset available is limited, incomplete, and even some target-aspects of HRRP are deficient. In these cases, generalization performance is an important criterion to evaluate a model.

We set the training dataset with half of the target-aspect angles in previous experiments. Without loss of generality, data from the 2nd segment of Yak-42, the 6th of Cessna Citation S/II and the 5th of An-26 are taken as the samples in the incomplete training dataset, and the test dataset is the same as previous experiments. Since the number of training samples is reduced by half, we cut down the number of hidden units in deep networks to avoid overfitting, i.e., $K_1 = 1300$, $K_2 = 200$ and $K_3 = 50$, respectively.

Table 4 depicts that the recognition accuracies of different models in the case of half training dataset. We can see that deep networks work better than the shallow models, among which SCAE still provides the best recognition performance. This indicates deep models robust to the incompleteness of training data. To further study the generalization performance, we create visualizations to understand the behavior and training of SCAE. A measured HRRP sequence from two other large or medium-sized civilian aircrafts is considered as confuser target data in the following experiments.

Fig. 14 shows the mapping of the high level features of SCAE to the first two principal components trained without the confuser target data, where deep models produce more discriminative feature space and learn the highly general representation of HRRP, even with the incomplete training dataset.

When the amount of data under test is huge, and the training data is limited, missing or non-observed, models may be trained to overfitting and tend to mismatch the unseen samples. In this experiment, we present the reconstruction performance of SCAE. Four types of experiment settings are considered: a training sample in the incomplete training set, a training sample out of the incomplete training set, a test sample and an HRRP in the confuser target dataset, respectively. Note that the last two (for complete training set) or three samples (for incomplete training set) are unseen in the training phase, which make the task not that easy. Fig. 15 shows that SCAE is able to retain the main details in HRRP and does not have the mismatch phenomenon. It indicates that SCAE has a good generalization capability to distribute the characteristics of HRRP even unseen.

In practice, traditional recognition methods generally rely on human ingenuity and require complete training samples to obtain favorable performance. These methods are unable to disentangle the intrinsic discriminative information and the factors of target-aspect variation. Therefore they do not work well even though there are still sufficient training samples for recognition task in finite target-aspects. But, "deep" architecture can deal with this case since its capability to disentangle as many factors as possible, discarding as little information about the data as is practical [18]. More generally, in deep networks, low-level features are shared among categories, while high-level features are more global and more invariant to most of the variations that are typically present in the training distribution (e.g. target-aspect variation). Therefore, the better generalization performance is an excellent advantage of SCAE, which brings the good engineering application prospect for HRRP-based RATR.

5. Conclusions

In this paper, deep networks are analyzed and utilized as a novel feature extraction method for HRRP-based RATR, and achieve better recognition performance than several traditional feature extraction methods in the experiments based on measured HRRP data. In order to learn more useful and stable hierarchical features for RATR from HRRP data, we further propose a novel deep architecture named SCAE. The proposed model employs the average profile as the correction terms in the networks, and the covariance matrix of each HRRP frame is also considered for establishing an effective loss function under the Mahalanobis distance criterion. The experimental results on measured HRRP data show that our model achieves superior recognition performance and generalization performance than several shallow or deep models based on feature extraction, even with an incomplete training set.

Acknowledgment

This work is partially supported by Program for Young Thousand Talent by Chinese Central Government, the National Natural Science Foundation of China (61372132, 61201292), the National Science Fund for Distinguished Young Scholars (61525105), Program for New Century Excellent Talents in University (NCET-13-0945), the National Defense Pre-research Fund (9140A07010115DZ01015, 9140A07010115DZ01019) and the Fundamental Research Funds for the Central Universities.

Appendix A. The conditional probabilities of GRBM

The energy function and the joint probability of GRBM are

$$E(\mathbf{x},\mathbf{h})=\sum_{i=1}^{D}\frac{(x_i-c_i)^2}{2\sigma_i^2}-\sum_{j=1}^{K}b_jh_j-\sum_{i=1}^{D}\sum_{j=1}^{K}\frac{x_i}{\sigma_i}h_jW_{ij} \tag{A1}$$

$$P(\mathbf{x},\mathbf{h})=\frac{1}{Z}\exp\big(-E(\mathbf{x},\mathbf{h})\big) \tag{A2}$$

where x_i denotes the real-valued visible unit and $Z=\sum_{\mathbf{x}}\sum_{\mathbf{h}}\exp\big(-E(\mathbf{x},\mathbf{h})\big)$. Each visible unit adds a parabola (quadratic) offset to the energy function, where σ_i controls the width of the parabola. The conditional probabilities of individual x_i and h_k can be explicitly written as follows:

$$\begin{aligned}
P(\mathbf{x}|\mathbf{h})&=\frac{P(\mathbf{x},\mathbf{h})}{P(\mathbf{h})}=\frac{e^{-E(\mathbf{x},\mathbf{h})}}{\int_{\mathbf{u}}e^{-E(\mathbf{u},\mathbf{h})}d\mathbf{u}}\\
&=\frac{e^{-\sum_{i=1}^{D}\frac{(x_i-c_i)^2}{2\sigma_i^2}+\sum_{j=1}^{K}b_jh_j+\sum_{i=1}^{D}\sum_{j=1}^{K}\frac{x_i}{\sigma_i}h_jW_{ij}}}{\int_{\mathbf{u}}e^{-\sum_{i=1}^{D}\frac{(u_i-c_i)^2}{2\sigma_i^2}+\sum_{j=1}^{K}b_jh_j+\sum_{i=1}^{D}\sum_{j=1}^{K}\frac{u_i}{\sigma_i}h_jW_{ij}}d\mathbf{u}}\\
&=\frac{e^{\sum_{j=1}^{K}b_jh_j}.e^{-\sum_{i=1}^{D}\frac{(x_i-c_i)^2}{2\sigma_i^2}+\sum_{i=1}^{D}\sum_{j=1}^{K}\frac{x_i}{\sigma_i}h_jW_{ij}}}{e^{\sum_{j=1}^{K}b_jh_j}.\int_{\mathbf{u}}e^{-\sum_{i=1}^{D}\frac{(u_i-c_i)^2}{2\sigma_i^2}+\sum_{i=1}^{D}\sum_{j=1}^{K}\frac{u_i}{\sigma_i}h_jW_{ij}}d\mathbf{u}}\\
&=\frac{e^{-\sum_{i=1}^{D}\frac{(x_i-c_i)^2}{2\sigma_i^2}+\sum_{i=1}^{D}\sum_{j=1}^{K}\frac{x_i}{\sigma_i}h_jW_{ij}}}{\prod_{i=1}^{D}\left[e^{\frac{1}{2}\left(\sum_{j=1}^{K}h_jW_{ij}\right)^2+\frac{1}{\sigma_i}c_i\sum_{j=1}^{K}h_jW_{ij}}.\sigma_i\sqrt{2\pi}\right]}\\
&=\frac{\prod_{i=1}^{D}e^{-\frac{(x_i-c_i)^2}{2\sigma_i^2}+\frac{x_i}{\sigma_i}\sum_{j=1}^{K}h_jW_{ij}}}{\prod_{i=1}^{D}\left[e^{\frac{1}{2}\left(\sum_{j=1}^{K}h_jW_{ij}\right)^2+\frac{1}{\sigma_i}c_i\sum_{j=1}^{K}h_jW_{ij}}.\sigma_i\sqrt{2\pi}\right]}=\\
&=\prod_{i=1}^{D}\frac{1}{\sigma_i\sqrt{2\pi}}.e^{-\frac{(x_i-c_i)^2}{2\sigma_i^2}-\frac{1}{2}\left(\sum_{j=1}^{K}h_jW_{ij}\right)^2+\frac{x_i-c_i}{\sigma_i}\sum_{j=1}^{K}h_jW_{ij}}\\
&=\prod_{i=1}^{D}\frac{1}{\sigma_i\sqrt{2\pi}}\\
&\quad .e^{-\frac{1}{2\sigma_i^2}\left[(x_i-c_i)^2+\sigma_i^2\left(\sum_{j=1}^{K}h_jW_{ij}\right)^2-2\sigma_i(x_i-c_i)\left(\sum_{j=1}^{K}h_jW_{ij}\right)\right]}\\
&=\prod_{i=1}^{D}\frac{1}{\sigma_i\sqrt{2\pi}}.e^{-\frac{1}{2\sigma_i^2}\left[x_i-c_i-\sigma_i\sum_{j=1}^{K}h_jW_{ij}\right]^2}
\end{aligned} \tag{A3}$$

Since there no connections between units in the same layer of the undirected graph (GRBM), visible units are conditionally independent given the hidden units (and vice versa).

$$p(x_i=x|\mathbf{h})=\mathcal{N}\left(x|c_i+\sigma_i\sum_j h_jW_{ij},\ \sigma_i^2\right) \tag{A4}$$

$$\begin{aligned}
p(h_k=1|\mathbf{x})&=\frac{\sum_{\mathbf{h}_{j\neq k}}p(\mathbf{x},\ h_k=1,\ \mathbf{h}_{j\neq k})}{p(\mathbf{x})}=\frac{\sum_{\mathbf{h}_{j\neq k}}e^{-E(\mathbf{x},h_k=1,\mathbf{h}_{j\neq k})}}{\sum_{\mathbf{g}}e^{-E(\mathbf{x},\mathbf{g})}}\\
&=\frac{\sum_{\mathbf{h}_{j\neq k}}e^{\left(b_k+\sum_{i=1}^{D}\frac{x_i}{\sigma_i}W_{ik}\right)+\left(\sum_{i=1}^{D}\sum_{j\neq k}^{K}\frac{x_i}{\sigma_i}h_jW_{ij}-\sum_{i=1}^{D}\frac{(x_i-c_i)^2}{2\sigma_i^2}+\sum_{j\neq k}^{K}b_jh_j\right)}}{\sum_{\mathbf{g}}e^{-E(\mathbf{x},\mathbf{g})}}\\
&=\frac{e^{\left(b_k+\sum_{i=1}^{D}\frac{x_i}{\sigma_i}W_{ik}\right)}\sum_{\mathbf{h}_{j\neq k}}e^{\left(\sum_{i=1}^{D}\sum_{j\neq k}^{K}\frac{x_i}{\sigma_i}h_jW_{ij}-\sum_{i=1}^{D}\frac{(x_i-c_i)^2}{2\sigma_i^2}+\sum_{j\neq k}^{K}b_jh_j\right)}}{\sum_{\mathbf{g}}e^{-E(\mathbf{x},\mathbf{g})}}\\
&=\frac{e^{\left(b_k+\sum_{i=1}^{D}\frac{x_i}{\sigma_i}W_{ik}\right)}\sum_{\mathbf{h}_{j\neq k}}e^{-E\left(\mathbf{x},h_k=0,\mathbf{h}_{j\neq k}\right)}}{\sum_{\mathbf{g}_{j\neq k}}e^{-E\left(\mathbf{x},g_k=0,\mathbf{g}_{j\neq k}\right)}+\sum_{\mathbf{g}_{j\neq k}}e^{-E\left(\mathbf{x},g_k=1,\mathbf{g}_{j\neq k}\right)}}\\
&=\frac{e^{\left(b_k+\sum_{i=1}^{D}\frac{x_i}{\sigma_i}W_{ik}\right)}\sum_{\mathbf{h}_{j\neq k}}e^{-E\left(\mathbf{x},h_k=0,\mathbf{h}_{j\neq k}\right)}}{\sum_{\mathbf{g}_{j\neq k}}e^{-E\left(\mathbf{x},g_k=0,\mathbf{g}_{j\neq k}\right)}+\sum_{\mathbf{g}_{j\neq k}}e^{-E\left(\mathbf{x},g_k=1,\mathbf{g}_{j\neq k}\right)}}\\
&=\frac{e^{\left(b_k+\sum_{i=1}^{D}\frac{x_i}{\sigma_i}W_{ik}\right)}\sum_{\mathbf{h}_{j\neq k}}e^{-E\left(\mathbf{x},h_k=0,\mathbf{h}_{j\neq k}\right)}}{\sum_{\mathbf{g}_{j\neq k}}e^{-E\left(\mathbf{x},g_k=0,\mathbf{g}_{j\neq k}\right)}+e^{\left(b_k+\sum_{i=1}^{D}\frac{x_i}{\sigma_i}W_{ik}\right)}\sum_{\mathbf{g}_{j\neq k}}e^{-E\left(\mathbf{x},g_k=0,\mathbf{g}_{j\neq k}\right)}}\\
&=\frac{e^{\left(b_k+\sum_{i=1}^{D}\frac{x_i}{\sigma_i}W_{ik}\right)}}{1+e^{\left(b_k+\sum_{i=1}^{D}\frac{x_i}{\sigma_i}W_{ik}\right)}}=\frac{1}{1+e^{-\left(b_k+\sum_{i=1}^{D}\frac{x_i}{\sigma_i}W_{ik}\right)}}\\
&=\mathrm{sigm}\left(b_k+\sum_{i=1}^{D}\frac{x_i}{\sigma_i}W_{ik}\right)
\end{aligned} \tag{A5}$$

References

[1] B. Chen, H.W. Liu, J. Chai, Z. Bao, Large margin feature weighting method via linear programming, IEEE Trans. Knowl. Data Eng. 21 (10) (2009) 1475–1488.

[2] L. Du, H.W. Liu, Z. Bao, J.Y. Zhang, Radar automatic target recognition using complex high-resolution range profiles, Proc. IET Radar Sonar Navig. 1 (1) (2007) 18–26.

[3] L. Du, H.W. Liu, P.H. Wang, B. Feng, M. Pan, F. Chen, Z. Bao, Noise robust radar HRRP target recognition based on multitask factor analysis with small training data size, IEEE Trans. Signal Process. 60 (7) (2012) 3546–3559.

[4] S.P. Jacobs, Automatic target recognition using high-resolution radar range profiles (Ph.D. dissertation), Washington Univ.,, St. Louis, MO, 1999.

[5] X.D. Zhang, Y. Shi, Z. Bao, A new feature vector using selected bispectra for signal classification with application in radar target recognition, IEEE Trans. Signal Process. 49 (9) (2001) 1875–1885.

[6] L. Du, H.W. Liu, Z. Bao, M.D. Xing, Radar HRRP target recognition based on higher-order spectra, IEEE Trans. Signal Process. 53 (7) (2005) 2359–2368.

[7] B. Feng, L. Du, H. W. Liu, et al., Radar HRRP target recognition based on K-SVD algorithm, in: Proceedings of the IEEE CIE International Conference on Radar, China, 2011, pp. 642–645.

[8] P. Molchanov, K. Egiazarian, J. Astola, et al., Classification of aircraft using micro-Doppler bicoherence-based features, IEEE Trans. Aerosp. Electron. Syst. 50 (2) (2014) 1455–1467.

[9] H.J. Li, S.H. Yang, Using range profiles as feature vectors to identify aerospace objects, IEEE Trans. Antennas Propag. 41 (3) (1993) 261–268.

[10] A. Zyweck, R.E. Bogner, Radar target classification of commercial aircraft, IEEE Trans. Aerosp. Electron. Syst. 32 (2) (1996) 598–606.

[11] X.J. Liao, P. Runkle, L. Carin, Identification of ground targets from sequential high-range-resolution radar signatures, IEEE Trans. Aerosp. Electron. Syst. 38 (4) (2002) 1230–1242.

[12] F. Zhu, X.D. Zhang, Y.F. Hu, D. Xie, Nonstationary hidden Markov models for multiaspect discriminative feature extraction from radar targets, IEEE Trans. Signal Process. 55 (5) (2007) 2203–2213.

[13] L. Du, H.W. Liu, Z. Bao, Radar HRRP statistical recognition: parametric model

and model selection, IEEE Trans. Signal Process. 56 (5) (2008) 1931–1944.
[14] L. Du, P.H. Wang, H.W. Liu, M. Pan, F. Chen, Z. Bao, Bayesian spatiotemporal multitask learning for radar HRRP target recognition, IEEE Trans. Signal Process. 59 (7) (2011) 3182–3196.
[15] M.D. Xing, Z. Bao, B. Pei, The properties of high-resolution range profiles, Opt. Eng. 41 (2) (2002) 493–504.
[16] L. Shi, P.H. Wang, H.W. Liu, et al., Radar HRRP statistical recognition with local factor analysis by automatic Bayesian Ying-Yang harmony learning, IEEE Trans. Signal Process. 59 (2) (2011) 610–617.
[17] B. Chen, H.W. Liu, Z. Bao, Analysis of Three Kinds of Classification Based on Different Absolute alignment Methods, Mod. Radar 28 (3) (2006) 58–62.
[18] Y. Bengio, A. Courville, P. Vincent, Representation learning: a review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1798–1828.
[19] Y. Bengio, Learning Deep Architectures for AI, Found. Trends Mach. Learn. 2 (1) (2009) 1–127.
[20] G. Hinton, R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (5786) (2006) 504–507.
[21] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.A. Manzagol, Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion, J. Mach. Learn. Res. 11 (2010) 3371–3408.
[22] A. Fischer, C. Igel, An Introduction to Restricted Boltzmann Machines, CIARP 2012, LNCS 7441, 2012, pp. 14–36.
[23] K. Jarrett, K. Kavukcuoglu, M. Ranzato, Y. LeCun, What is the best multi-stage architecture for object recognition? in: Proceedings of IEEE Int. Conf. Comp. Vision (ICCV), 2009.
[24] H. Lee, C. Ekanadham, A.Y. Ng, Sparse deep belief network model for visual area V2, in: Proc. Neural Info. Processing Systems (NIPS), 2007.
[25] J. Y. Xie, L. L. Xu and E. H. Chen, Image Denoising and Inpainting with Deep Neural Networks, Proc. Neural Info. Processing Systems (NIPS), 2012.
[26] B. Chen, G. Polatkan, G. Sapiro, et al., Deep learning with hierarchical convolutional factor analysis, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1887–1901.
[27] G. Hinton, L. Deng, G.E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, B. Kingsbury, Deep neural networks for acoustic modeling in speech recognition, IEEE Signal Process. Mag. 29 (6) (2012) 82–97.
[28] Y. Bengio, Deep Learning of Representations for Unsupervised and Transfer Learning, JMLR Workshops and Conf. Proc., 27, 2012, pp. 17–36.
[29] A. Krizhevsky, G. Hinton, Learning Multiple Layers of Features from Tiny Images, Technical report, Univ. of Toronto, 2009.
[30] H. Yu, J. Yang, A direct LDA algorithm for high-dimensional data with application to face recognition, Pattern Recognit. 34 (10) (2001) 2067–2070.
[31] Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle, Greedy layer-wise training of deep networks, in: Proc. Neural Information and Processing Systems, 2007.
[32] V. Nair, G.E. Hinton, 3D object recognition with deep belief nets, in: Proc. Neural Information and Processing Systems, 2009.
[33] G.E. Hinton, A Practical Guide to Training Restricted Boltzmann Machines, Technical Report UTML TR 2010-003, Dept. of Computer Science, Univ. of Toronto, 2010.
[34] Y. Bengio, Practical Recommendations for Gradient-Based Training of Deep Architectures, Neural Networks: Tricks of the Trade, Springer, 2013.
[35] D.L. Donoho, M. Elad, V. Temlyakov, Stable recovery of sparse overcomplete representations in the presence of noise, IEEE Trans. Inf. Theory 52 (1) (2006) 6–18.

Bo Feng received the BEng degree in electronic engineering from Xidian University in 2009 and his Ph.D. degree from National Lab of Radar Signal Processing, Xidian University in 2015. His research interests include machine learning, radar signal processing, and radar automatic target recognition.

Bo Chen received the B.S., M.S., and Ph.D. degrees from Xidian University, Xi'an, China, in 2003, 2006, and 2008, respectively, all in electronic engineering. He became a Post-Doctoral Fellow, a Research Scientist, and a Senior Research Scientist with the Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA, from 2008 to 2012. From 2013, he has been a Professor with the National Laboratory for Radar Signal Processing, Xidian University. He received the Honorable Mention for 2010 National Excellent Doctoral Dissertation Award and is selected into One Thousand Young Talent Program in 2013. His current research interests include statistical machine learning, statistical signal processing and radar automatic target detection and recognition.

Hongwei Liu received the M.S. and Ph.D. degrees in electronic engineering from Xidian University in 1995 and 1999, respectively. He is currently a Professor and the director of the National Laboratory of Radar Signal Processing, Xidian University, Xi'an. He worked at the National Laboratory of Radar Signal Processing, Xidian University, Xi'an. From 2001 to 2002, he was a Visiting Scholar at the Department of Electrical and Computer Engineering, Duke University, Durham, NC. He is currently a Professor and the director of the National Laboratory of Radar Signal Processing, Xidian University, Xi'an. His research interests are radar automatic target recognition, radar signal processing, adaptive signal processing, and cognitive radar.