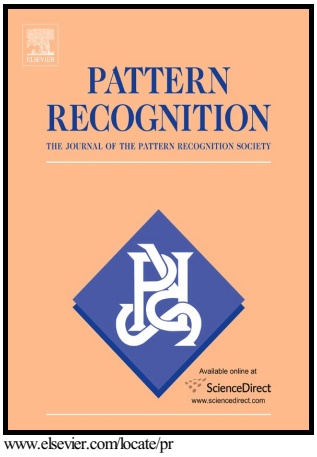
SAR Image Segmentation Based on Convolutional-wavelet Neural Network and Markov Random Field

Yiping Duan, Fang Liu, Licheng Jiao, Peng Zhao, Lu Zhang



www.elsevier.com/locate/pr

PII: S0031-3203(16)30372-7
DOI: http://dx.doi.org/10.1016/j.patcog.2016.11.015
Reference: PR5958

To appear in: *Pattern Recognition*

Received date: 27 March 2016
Revised date: 5 October 2016
Accepted date: 16 November 2016

Cite this article as: Yiping Duan, Fang Liu, Licheng Jiao, Peng Zhao and Lu Zhang, SAR Image Segmentation Based on Convolutional-wavelet Neural Network and Markov Random Field, *Pattern Recognition* http://dx.doi.org/10.1016/j.patcog.2016.11.015

SAR Image Segmentation Based on Convolutional-wavelet Neural Network and Markov Random Field

Yiping Duan[a,b], Fang Liu[a,b], Licheng Jiao[b], Peng Zhao[a,b], Lu Zhang[b]

[a]*School of Computer Science and Technology, Xidian University, Xi'an,710071, P.R. China*
[b]*Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, International Research Center for Intelligent Perception and Computation, Joint International Research Laboratory of Intelligent Perception and Computation, Xidian University, Xi'an, Shaanxi Province, 710071, China*
f63liu@163.com

Abstract— Synthetic aperture radar (SAR) imaging system is usually an observation of the earths' surface. It means that rich structures exist in SAR images. Convolutional neural network (CNN) is good at learning features from raw data automatically, especially the structural features. Inspired by these, we propose a novel SAR image segmentation method based on convolutional-wavelet neural networks (CWNN) and Markov Random Field (MRF). In this approach, a wavelet constrained pooling layer is designed to replace the conventional pooling in CNN. The new architecture can suppress the noise and is better at keeping the structures of the learned features, which are crucial to the segmentation tasks. CWNN produces the segmentation map by patch-by-patch scanning. The segmentation result of CWNN will be used with two labeling strategies (i.e., a superpixel approach and a MRF approach) to produce the final segmentation map. The superpixel approach is used to enforce the smooth nature on the local region. On the other hand, the MRF approach is used to preserve the edges and the details of the SAR image. Specifically, two segmentation maps will be produced by applying the superpixel and MRF approaches. The first segmentation map is obtained by combining the segmentation map of CWNN and the superpixel approach, and the second segmentation map is obtained by applying the MRF approach on the original SAR image. Afterwards, these two segmentation maps are fused by using the sketch map of the SAR image to produce the final segmentation map. Experiments on the texture images demonstrate that the CWNN is effective for the segmentation tasks. Moreover, the experiments on the real SAR images show that our approach obtains the regions with labeling consistency and preserves the edges and details at the same time.

Keywords： Convolutional Neural Network, wavelet transform, Markov Random Filed, SAR image segmentation.

1. INTRODUCTION

SYNTHETIC aperture radar (SAR) is an active imaging system with microwave which can be operated day and night under all weather conditions. It has been widely used in many practical applications, such as environment, traffic, civil applications, military applications and homeland security [1, 2]. The purpose of SAR image segmentation is to divide the image into regions of different characteristics without intersection, which consists of labeling every pixel in the image [3]. The segmentation is basic and important for the understanding and interpretation of SAR images [4, 5].

SAR image segmentation has been conducted with a wide variety of methods. They can be mainly divided into two categories. Firstly, the feature based approaches. Many researches pay much attention to extract SAR image features which include the texture features [6-11], edge features [12, 13] and hybrid features [4, 14]. Then, the features are used to train the classifier for labeling the image. The general workflow is shown in Fig. 1(a). Secondly, the statistical model based approaches [15-20]. They have the solid mathematical foundation and the ability to take the contextual information of the image into account. These methods mainly consist of a likelihood model and a prior model. The likelihood model describes the statistical property of the observed SAR image conditioned on the class labels and it is usually assumed to be Nakagami [21], Gamma [21], K [22], or G [22] distributions and so on. The prior model describes the interactions between the class labels, such as Gibbs distribution [18] and multinomial logistic function [20]. According to Bayes' rule, the posterior probability can be derived as the product of the likelihood distribution and the prior distribution. The class labels of the image are estimated by maximum a posteriori (MAP). The general workflow is shown in Fig. 1(b).

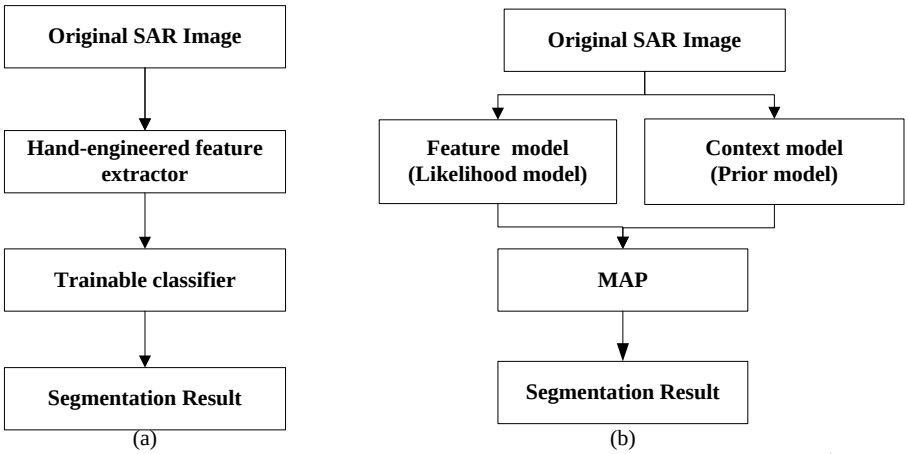


Fig. 1. Workflow of the typical approaches. (a) Workflow of the feature based approaches; (b) Workflow of the statistical model based approaches.

All the above methods are based on the hand-engineered features, which always gather the specified features from the input samples. But these features might not be suitable for the problem at hand. Neural networks provide a way to learn features from the raw pixels automatically. Convolutional neural network (CNN) is a popular type of neural network and designed specifically to recognize two dimensional (2D) shapes [23]. In CNN, the convolutional layer alternates with the pooling layer to learn the features, which have the invariance of translation, scaling, shift, and other forms of distortions [23]. Moreover, the shared weights in CNN reduce the free parameters, which ease the burden of the computing. Hence, CNN is widely used in image processing. They are best known for their applications in image classification and recognition [24-29]. This type of task is to predict one label of each image. By using CNN, each image is presented by the learned features. Then, the images are classified or recognized according to the learned features. These successful applications have shown that CNN obtains remarkably better results than the conventional methods. Moreover, they have also been used in image segmentation, particularly for natural images [30-34]. The segmentation task is to predict one label for each pixel and have some differences with the whole-image classification. The input samples in the segmentation task are image patches surrounding the central pixels. Since the segmentation method with image patch can capture

contextual information, it will bring good segmentation accuracy and labeling consistency. The larger the patch is, the better labeling consistency can be achieved [30]. However, when the image patch is very large, some details will be lost and the edges will be obscure. Therefore, in order to overcome the drawbacks of the segmentation method with image patch, we fuse the segmentation results of CWNN and Markov random field (MRF) model together.

CNN has achieved continuously promising results in image segmentation. However, conventional pooling (max pooling) in CNN does not take into account the structure of the previous layer, blindly taking the maximum value in the rectangular window as the output. This may lead to the loss of some structures (such as edges, endpoints, corners). A more convincing way is to produce the feature map according to the specified rules. Wavelet transform [35, 36] is an effective way to satisfy the condition. It can better match the human visual system (HVS) and extracts the features according to the specified rules. The wavelet coefficients in the low frequency subbands represent the average brightness of the image, where the large coefficients represent the region with high average brightness and the small coefficients represent the region with low average brightness. The wavelet coefficients in the high frequency subbands represent the textures and the edges of the image, where the large coefficients represent the complex texture and the edges, and the small coefficients represent the smooth part of the image. It indicates that the wavelet transform considers the structures of the image while extracting the features. Hence, the wavelet transform is plugged into CNN to improve the network performance.

Motivated by the discussions above, we propose a novel SAR image segmentation method based on convolutional-wavelet neural network (CWNN) and MRF. In the approach, we propose a new network architecture named CWNN by replacing the conventional pooling layer with a wavelet constrained pooling layer. In CWNN, the convolutional layer alternates with the wavelet pooling layer to learn the features. The patches surrounding the central pixels are fed into CWNN to predict

the labels of the centered pixels by the forward propagation and the parameters are learned by standard back propagation. The segmentation result of CWNN will be used with two labeling strategies (i.e., a superpixel approach and a MRF approach) to produce the final segmentation map. The superpixel approach is used to enforce the smooth nature on the local region. On the other hand, the MRF approach is used to preserve the edges and the details of the SAR image. Specifically, two segmentation maps will be produced by applying the superpixel approach and the MRF approach. The first segmentation map is obtained by combining the segmentation map of CWNN and the superpixel approach, and the second segmentation map is obtained by applying the MRF approach on the original SAR image. Afterwards, these two segmentation maps are fused by using the sketch map [37] of the SAR image to produce the final segmentation map. The workflow of the proposed approach is depicted in Fig. 2. It should be noted that only the forward propagation is shown in Fig. 2. The main contributions of our approach can be summarized as follows, 1) a wavelet constrained pooling layer is designed to replace the conventional pooling. The wavelet pooling extracts the features according to the specified rules; 2) the wavelet pooling is plugged into the original CNN to generate the new network architecture named CWNN. The new network architecture can suppress the noise and is better at keeping the structures of the learned features. This contributes a more reliable segmentation; 3) two labeling strategies are used to refine the segmentation results. Specifically, the labeling strategy based on the superpixel approach is used to force the local region into the same label. The labeling strategy based on the sketch map [37] is used to locate the edges and details accurately.

The rest of this paper is organized as follows. In section 2, we give a brief overview on CNN. Then, with the potential improvements, a new network architecture CWNN is proposed. In section 3, we describe the SAR image segmentation based on the MRF approach. In section 4, the labeling strategies, which are used in refining the result of CWNN are presented. In section 5, the whole

segmentation process is introduced step by step. Experimental results and analyses are presented in section 6. Section 7 concludes this paper and presents some perspectives for our future work.

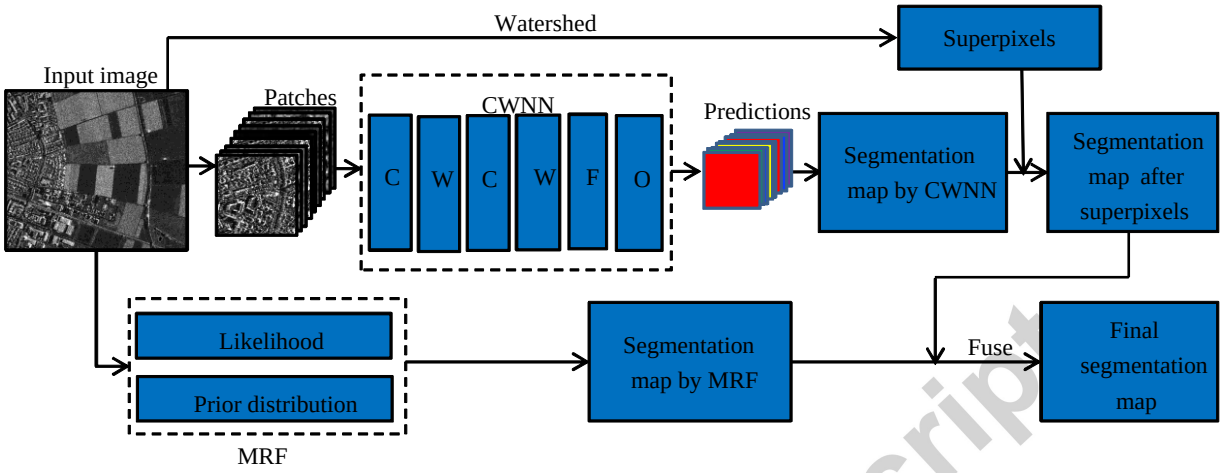


Fig. 2. The flowchart of our approach.

2. Convolutional-wavelet neural network

In this section, we give a brief review and some conclusions about CNN. Then, a new network architecture named CWNN is proposed with some improvements based on the original CNN.

2.1 The analysis of convolutional neural network

CNN is designed to deal with the variability of the 2D shapes [23]. They have been extensively investigated to extract features for image processing. A basic stage in CNN is composed of a convolutional layer and a pooling layer. In general, a CNN is composed of multiple stages, followed by a classification module. The inputs and the outputs of the convolutional layer and the pooling layer are called feature maps. Through the propagation of CNN, the sizes of the feature maps decrease layer by layer and the extracted features are more and more abstract and global. The details will be introduced below according to [38].

Given an image $\mathbf{I}$. The input image patches are represented as $\mathbf{x}_0 = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_S\}$, where S is the number of the training samples. These input samples are obtained by extracting the patches centered

on the pixels $o=\{o_1,o_2,...,o_S\}$. For the input image patches $\mathbf{x}_0=\{\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_S\}$, the corresponding network output is expressed as $y=\{y_1,y_2,...,y_S\}$. Each y_s takes its value from a finite set of classes $\Omega_1=\{1,2,...,K\}$, where K is the number of class. l is the level of the network and it takes its value from a finite set of levels $\Omega_2=\{1,2,...,L\}$. The convolutional layer convolves the output of the preceding layer (input or pooling layer) with a sliding filter bank to generate the output feature maps. It is expressed as

$$g_j^l = sig\left(\sum_{i=1}^{m} x_i^{l-1} * w_{ji}^l + b_j^l\right) \tag{1}$$

where g_j^l is the j th output feature map in the l layer, and x_i^{l-1} is the i th input feature map in the $(l-1)$ th layer. The function $sig(\cdot)$ is a sigmoid function and used as the activation function in the network. The filter w_{ji}^l and the bias b_j^l constitute the trainable parameters of the convolutional layer.

The pooling layer reduces the resolution of the feature maps and reduces the sensitivity of the output to shifts. Max pooling is most commonly used in CNN. Specifically, the max pooling is written as

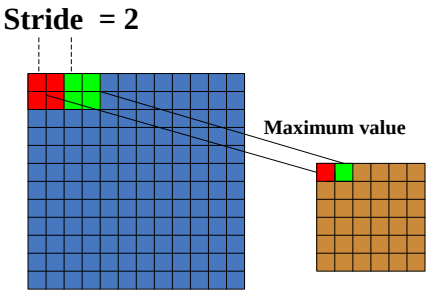
$$c_j^l = \max\{in_1, in_2, in_3, in_4\} \tag{2}$$

$in_1 = x_i^{l-1}\left(1:r:size(x_i^{l-1},1),1:r:size(x_i^{l-1},2)\right)$, $in_2 = x_i^{l-1}\left(1:r:size(x_i^{l-1},1),2:r:size(x_i^{l-1},2)\right)$,

$in_3 = x_i^{l-1}\left(2:r:size(x_i^{l-1},1),1:r:size(x_i^{l-1},2)\right)$, $in_4 = x_i^{l-1}\left(2:r:size(x_i^{l-1},1),2:r:size(x_i^{l-1},2)\right)$.

where c_j^l is the j th output feature maps of the pooling layer, and x_i^{l-1} is the i th input feature map in the $(l-1)$ th layer. The stride of the pooling layer r is set as 2. Moreover, the first dimension size and the second dimensional size of x_i^{l-1} are represented as $size\left(x_i^{l-1},1\right)$ and $size\left(x_i^{l-1},2\right)$, respectively In Fig. 3, we give an example of the max pooling. In the red block in Fig. 3(a), the

maximum value is selected as the output in Fig. 3(b).



(a) (b)

Fig. 3. An example of max pooling. (a) The input feature map of the pooling layer; (b) The output feature map by the max pooling.

Following the convolutional layer and the pooling layer is a full connected layer. The input feature map is converted into a vector $\mathbf{x}^{l-1}$. The dimension of the vector is n. The output of the full connected layer is

$$a_j^l = sig\left(\sum_{i=1}^{n} x_i^{l-1} \times w_{ji}^l + b_j^l\right) \tag{3}$$

where a_j^l is the j th output unit of the full connected layer, and x_i^{l-1} is the i th element in $\mathbf{x}^{l-1}$. The filter w_{ji}^l and the bias b_j^l constitute the trainable parameters of the full connected layer.

Finally, the output layer is composed of softmax units [39]. The output is computed as follows

$$p\left(y_s = j \mid \mathbf{a}_s; \boldsymbol{\theta}\right) = \frac{e^{\theta_j^{\mathrm{T}} \mathbf{a}_s}}{\sum_{j=1}^{K} e^{\theta_j^{\mathrm{T}} \mathbf{a}_s}} \tag{4}$$

where y_s is the actual output of the network and K is the number of the class. $\mathbf{a}_s$ is the feature vector derived by the full connected layer and $\boldsymbol{\theta}$ is the parameter of the softmax function.

The convolutional layer, pooling layer, full-connected layer and softmax classifier are the process of the forward propagation, which is used in predicting the labels of the input patches. The backward propagation (BP) [40] is the process of learning the trainable parameters. The loss

function in our approach is expressed as in [39],

$$E(\theta) = -\frac{1}{S}\left[\sum_{s=1}^{S}\sum_{j=1}^{K} 1\{y_s = j\} \log \frac{e^{\theta_j^{\mathrm{T}} \mathbf{a}_s}}{\sum_{j=1}^{K} e^{\theta_j^{\mathrm{T}} \mathbf{a}_s}}\right] + \frac{\lambda}{2}\sum_{i=1}^{K}\sum_{j=1}^{V}\theta_{ij}^2 \tag{5}$$

where the first term is the error item and the second term is the weight decay item. $1\{y_s = j\} = 1$ when $y_s = j$; $1\{y_s = j\} = 0$ when $y_s \neq j$. y_s is the actual output of the network. S is the number of the training samples, while V is the dimension of the feature vector. The trainable parameter θ_{ij} is obtained by the learning method. The gradient descent algorithm is used to learn the parameters $\mathbf{w}$, $\mathbf{b}$ and $\boldsymbol{\theta}$. The details of such an implementation are presented in [23] and [40].

2.2. Convolutional-wavelet neural network

CNN is mostly used in whole-image classification, i.e., predicting one label for a whole image. It has achieved good performance in many databases of natural images. The aim of CNN is to learn the discriminative features. It means that CNN is to learn the features, which can distinguish one class from another. At the beginning layer, CNN extracts elementary visual features such as oriented edges, endpoints, corners. These features are then combined by the subsequent layers in order to detect higher features [23]. However, the noise will affect CNN to learn the features. Hence, it is necessary to remove the noise. In addition, the conventional pooling (max pooling) in CNN do not take into account the structure of the previous layer, blindly taking the maximum value in the rectangular window as the output. Some structures (such as edges, endpoints, corners) may be lost by the simple pooling. It is necessary to try to preserve more structures (features) at the pooling layer. Furthermore, shift will cause the position of salient features to vary [23]. So, the shift invariance is important for CNN. Dual tree complex wavelet transform (DT-CWT) can remove the

noise and preserve the structure (edges, endpoints, corners) of the image at the same time. Moreover, DT-CWT has five advantages, such as approximate shift invariance, good directional selectivity, perfect reconstruction, limited redundancy and efficient order-N computation [41-44]. Therefore, we plug DT-CWT into the original CNN to generate the new network architecture. By using DT-CWT, the preceding layer is decomposed into eight components, two low frequency subbands LL_1 and LL_2, and the high frequency subbands in six orientations, $\pm 15^{\circ}, \pm 45^{\circ}$ and $\pm 75^{\circ}$ which are written as LH_1, LH_2, HL_1, HL_2, HH_1 and HH_2 [41-44]. The average of the two low frequency subbands is selected as the output of the pooling layer. On one hand, low-frequency subbands keep the structures of input layer according to the specified rules, which lead to a better representation for the input image patch. On the other hand, by losing the high frequency subbands, some noises can be suppressed. We use one-level DT-CWT decomposition in our method. Similar to max-pooling, the input of a wavelet pooling layer is the output of the previous convolutional layer. For each input feature map x_i^{l-1}, we use DT-CWT to obtain the subbands

$$\left\{\mathrm{LL}_1, \mathrm{LL}_2, \mathrm{LH}_1, \mathrm{LH}_2, \mathrm{HL}_1, \mathrm{HL}_2, \mathrm{HH}_1, \mathrm{HH}_2\right\} = f\left(x_i^{l-1}\right) \tag{6}$$

where the dual-tree complex wavelet transform $f(\cdot)$ produces the eight components $\mathrm{LL}_1, \mathrm{LL}_2, \mathrm{LH}_1$, $\mathrm{LH}_2, \mathrm{HL}_1, \mathrm{HL}_2, \mathrm{HH}_1, \mathrm{HH}_2$.

Then the average of the low-frequency subbands is used as the output of the wavelet pooling layer, which is defined as

$$\hat{c}_j^l = \frac{1}{2}\left(\mathrm{LL}_1 + \mathrm{LL}_2\right) \tag{7}$$

where $\hat{c}_j^l$ is the j th output of the wavelet pooling layer.

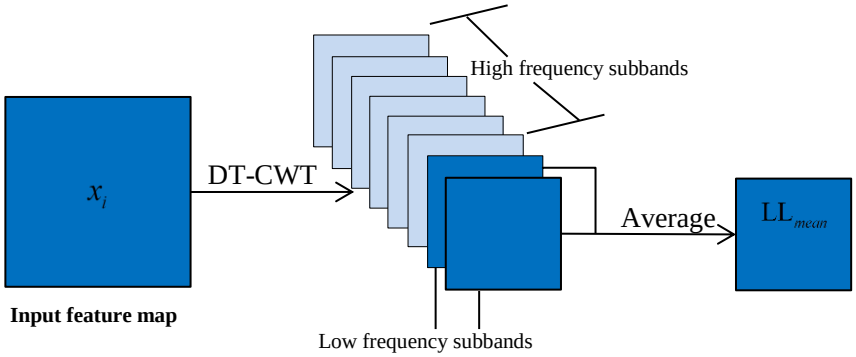


Fig. 4.Wavelet pooling layer.

In Fig. 4, we give an example to show the wavelet pooling layer in Fig. 4. x_i is one feature map after the convolutional layer. DT-CWT is carried out on the feature map x_i to produce eight subbands, two approximation subbands (in blue) and six high frequency subbands (in gray). The two approximation subbands are averaged to obtain the output feature map.

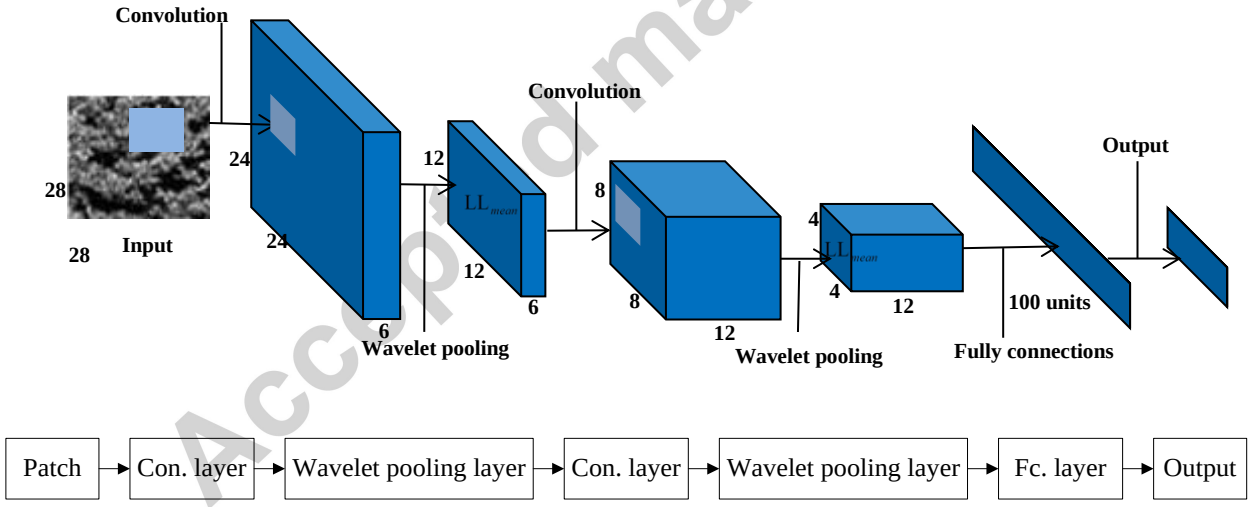


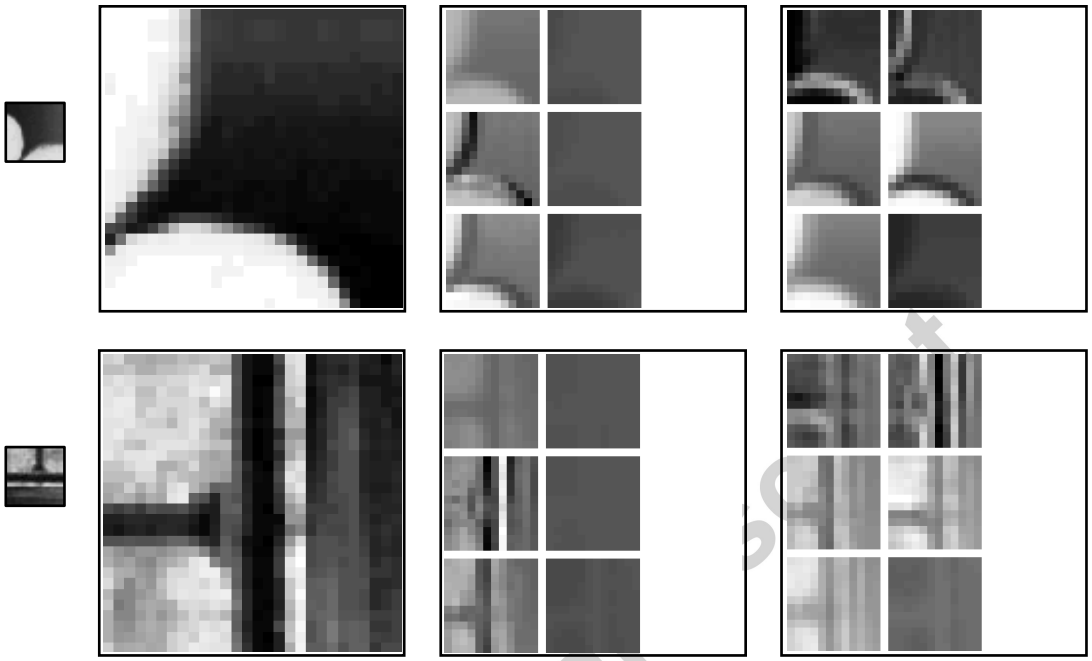
Fig. 5. The network architecture of CWNN.

The wavelet pooling is incorporated into the original CNN to generate the new network architecture CWNN (shown in Fig. 5). The convolutional layer is expressed as C_l, the wavelet pooling layer is expressed as W_l, where l is the index of the layer. The network is expressed as $\{I_1, C_2, W_3, C_4, W_5, F_6, O_7\}$. I_1 is the input layer with the patch size of 28×28. C_2 is a convolutional

layer with 6 convolutional kernels whose sizes are 5×5. This layer produces 6 feature maps with the size 24×24. W_3 is a wavelet pooling layer. In this layer, all the input feature maps are decomposed with one-level DT-CWT. LL_{mean} is used as the input of the next layer. This wavelet pooling layer produces 6 feature maps with the size of 12×12. C_4 is a convolutional layer with 12 convolutional kernels whose sizes are 5×5. This layer produces 12 feature maps with the size of 8×8. W_5 is a wavelet pooling layer. It produces 12 feature maps with the size of 4×4. F_6 is the full connected layer with 100 units. O_7 is the output layer with K units, where K is the number of the class. With this architecture, the convolutional layer alternates with wavelet pooling layer to learn the features.

In fact, the purpose of using the low frequency subbands of DT-CWT is to remove the noise and simultaneously preserve the structures of the feature map very well. The noise in the image will affect the features learned by CNN. At the same time, we want to preserve the structures of the feature maps. These structures are very important for the segmentation task. In order to better understand CWNN, we show the feature maps of different types of patches in Fig. 6. Fig. 6(a) is the original image patches with obvious structures. Fig. 6(b) is the visualization results of the original image patches. Fig. 6(c) and (d) are the feature maps after pooling layer in CNN and CWNN, respectively. It is observed that CNN and CWNN can learn the features very well. The feature maps learned from different patches have distinct differences. Moreover, we compare the feature maps learned by CNN in Fig. 6(c) and CWNN in Fig. 6(d). It is noteworthy that the features learned by CWNN maintain more structures than CNN. It shows that the features learned by CWNN represent the input image patch better than the features learned by CNN. It means that the wavelet pooling improves the performance of the network. The reason is that the wavelet pooling layer can not only

remove some noise but it can also keep the structures of the feature maps in lined with the specified rules.



(a) (b) (c) (d)

Fig. 6. Feature maps of different patches. (a) The original patches with the size of 28×28; (b) Visualization of the original patches; (c) The feature maps after max pooling in CNN; (d) The feature maps after wavelet pooling in CWNN.

3. MRF for SAR image segmentation

The MRF approach is widely used in SAR image segmentation [15-18]. It formulates the image segmentation into a probabilistic way. It is a pixel-based segmentation method and has a good ability to locate the edges and preserve the details. In the MRF approach, the segmentation problem is transformed into maximize the posterior (i.e., $p\left(y_{\hat{s}} \mid o_{\hat{s}}\right)$) problem, where $\hat{s}=\left\{1,2,...,\hat{S}\right\}$ is the sites of the image, $O=\left\{o_{\hat{s}} \mid \hat{s}\in\hat{S}\right\}$ is the observed field, $Y=\left\{y_{\hat{s}} \mid \hat{s}\in\hat{S}\right\}$ is the labeled field, each $y_{\hat{s}}$ takes its values in a finite set of classes $\Omega_1=\left\{1,2,...,K\right\}$. According to Bayes' rule, we have

$$y_{\hat{s}}=\arg\max_{y_{\hat{s}}\in\Omega_1}\left\{p\left(o_{\hat{s}} \mid y_{\hat{s}}\right)p\left(y_{\hat{s}}\right)\right\} \tag{8}$$

where the likelihood distribution $p(o_{\hat{s}} \mid y_{\hat{s}})$ represents the feature of the observed image and the prior distribution $p(y_{\hat{s}})$ represents the contextual information of the image.

The prior distribution $p(y_{\hat{s}})$ is used to describe the local spatial interactions between the labels. It is defined as a Gibbs distribution as in [18]

$$p(y_{\hat{s}}) = Z^{-1} \exp\left(-u(y_{\hat{s}})\right) \tag{9}$$

where Z is the normalized constant, and the function $u(y_{\hat{s}})$ is the clique energy function. In order to express the prior distribution clearly, the equation (9) is written as

$$p(y_{\hat{s}}) = \frac{\exp\left(-u(y_{\hat{s}})\right)}{\sum_{y_{\hat{s}} \in \Omega_1} \exp\left(-u(y_{\hat{s}})\right)} \tag{10}$$

where $u(y_{\hat{s}}) = -\beta \sum_{i \in N_{\hat{s}}} \left[\delta(y_{\hat{s}}, y_i) - 1\right]$ and $\delta(y_{\hat{s}}, y_i) = 1$ when $y_{\hat{s}} = y_i$, otherwise $\delta(y_{\hat{s}}, y_i) = 0$. $N_{\hat{s}}$ is the neighborhood of $o_{\hat{s}}$.

The likelihood model $p(o_{\hat{s}} \mid y_{\hat{s}})$ is often assumed as a Nakagami distribution [20] for simplicity. It is formulated as:

$$p(o_{\hat{s}} \mid y_{\hat{s}}; \alpha_k, \beta_k) = \frac{2}{\Gamma(\beta_k)} \left(\frac{\beta_k}{\alpha_k}\right)^{\beta_k} o_{\hat{s}}^{2\beta_k - 1} e^{\left(-\beta_k \frac{o_{\hat{s}}^2}{\alpha_k}\right)} \tag{11}$$

where α_k, β_k are the parameters of the distribution. Their analytical expressions are given in [20].

The MAP criterion [18] is used to estimate the labels according to (8). Utilizing the local dependence among the pixels, MAP iteratively refines the labels based on the provisional estimate of those labels.

4. Labeling strategies

In this section, we give two labeling strategies to refine the segmentation result of CWNN. The

first strategy is to use a superpixel approach to obtain the local smooth output. The second strategy is to use the sketch map [37] of the SAR image to fuse the segmentation results of CWNN and the MRF approach together, which can preserve the edges and the details.

In SAR image segmentation, the labels of the pixels are strongly dependent on the output space of the network. In order to obtain the local smooth output, we use a superpixel approach to force the local region into the same label. The watershed method is used to obtain the superpixels [45]. For each superpixel, the histogram of the labels is computed. Then, the superpixel is marked as the corresponding label of the histogram peak. This label strategy is shown in Fig. 7. The segmentation map produced by CWNN is marked as $\mathbf{Y}_{CWNN}$. After the superpixel strategy is done, the segmentation map is marked as $\hat{\mathbf{Y}}_{CWNN}$.

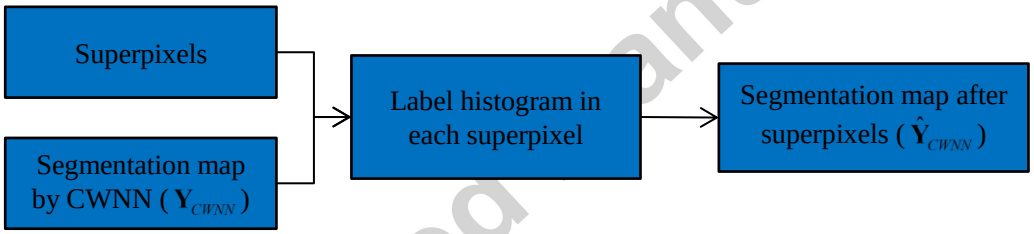


Fig. 7. Labeling strategy based on the superpixel approach.

CWNNs can produce abstract features and provide good performance in image segmentation. However, the abstract features hamper the accuracy of the location of the edges and the details. The local smoothness obtained using the superpixel strategy does not refine the edge and details. Hence, we fuse the results of CWNNs after the superpixel approach ($\hat{\mathbf{Y}}_{CWNN}$) and the segmentation result of the MRF approach ($\mathbf{Y}_{MRF}$) to remedy this problem. We need to find the true edge region, then the segmentation results of the true edge region is replaced by the segmentation result of the MRF approach ($\mathbf{Y}_{MRF}$). Here, we use sketch map [37] to find the true edge region. For a SAR image, we have the labeled part and the unlabeled part. The edges and details are included in the unlabeled

part. We compute the sketch map by using the edge-line intensity map [37]. The sketch map consists of sketch lines, which is a kind of sparse representation [46] of the SAR image and represents the positions of the gray changes [37, 47, 48]. According to the ground truth segmentation and the sketch map, we can get the sketch line in the unlabeled part. Then, the geometrical structure window [37, 47, 48] is operated on the sketch lines to obtain the region of edges and details. The size of the geometrical structure window is $5\times q$, where q is the length of the sketch line. Finally, the regions of edge and details are labeled by the result of the MRF approach ($\mathbf{Y}_{MRF}$) and the rest regions are labeled by the result of CWNN after the superpixel approach ($\hat{\mathbf{Y}}_{CWNN}$). By using this strategy, we can obtain the regions with labeling consistency and detail preservations in the segmentation map simultaneously. An example is given to show the fusion strategy in Fig. 8. Fig. 8(a) is an original SAR image. Fig. 8(b) is the corresponding ground truth segmentation. The white part is the unlabeled part and the rest is the labeled part. Fig. 8(c) is the edge-line intensity map [37]. Fig. 8(d) is the sketch map according to the edge-line intensity map. The white region in Fig. 8(e) is the region of edge and details. The region (white region) is labeled by the result of the MRF approach ($\mathbf{Y}_{MRF}$) and the rest region (black region) in Fig. 8(e) is labeled by the result of CNNs after the superpixel approach ($\hat{\mathbf{Y}}_{CWNN}$).

(a) (b) (c) (d) (e)

Fig. 8. The fusion strategy of $\hat{\mathbf{Y}}_{CWNN}$ and $\mathbf{Y}_{MRF}$. (a) Original SAR image (Noerdlinger Ries); (b) Ground truth segmentation; (c) Edge-line intensity map; (d) Sketch map; (e) A binary map, the pixels in the white portion are labeled by the MRF approach ($\mathbf{Y}_{MRF}$), and the pixels in the black portion are labeled by $\hat{\mathbf{Y}}_{CWNN}$.

5. Segmentation using the proposed approach

The overall segmentation process of our proposed approach is depicted in this section, which is composed of three parts: CWNN, a superpixel approach and a MRF approach. The architecture and parameters of CWNN are shown in Fig. 5. In the beginning, the training patches are fed into CWNN to train the network. Then, the testing patches are fed into the trained CWNN to predict the labels of the image and the segmentation map is marked as $\mathbf{Y}_{CWNN}$. Meanwhile, the watershed method and MRF model are operated on the original image to produce the superpixels and another segmentation map, respectively. Finally, the superpxiels and $\mathbf{Y}_{CWNN}$ are combined to improve the local smoothness and this segmentation map is marked as $\hat{\mathbf{Y}}_{CWNN}$. Afterwards, the segmentation map of the MRF approach and $\hat{\mathbf{Y}}_{CWNN}$ are fused to produce the final segmentation map. The main steps of our proposed approach are shown in Table 1.

Table 1
Algorithm description.

SAR image segmentation based on convolutional-wavelet neural network and Markov Random Field	
Data Pre-processing:	For each pixel $o=\{o_1,o_2,...,o_S\}$, extract the corresponding patches $\mathbf{x}_0=\{\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_S\}$ as the training samples and all the patches of the image $\mathbf{x}=\{\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_{\hat{S}}\}$ are used as the testing samples.

CWNN Forward Propagation:	For $l=\{1,2,...,L\}$ do
	If layer l ==convolutional layer then
	Convolute $\mathbf{w}_j$ with the feature map $g_j^l=sig\left(\sum_{i=1}^{m}x_i^{l-1}*w_{ji}^l+b_j^l\right)$ in *eq.(1)*
	Else if layer l ==wavelet pooling layer then
	Decompose each feature map with one-level DT-CWT
	$\{\mathrm{LL}_1,\mathrm{LL}_2,\mathrm{LH}_1,\mathrm{LH}_2,\mathrm{HL}_1,\mathrm{HL}_2,\mathrm{HH}_1,\mathrm{HH}_2\}=f\left(x_i^{l-1}\right)$ in *eq. (6)*
	Take the low-frequency subband as the output $\hat{c}_j^l=\frac{1}{2}(\mathrm{LL}_1+\mathrm{LL}_2)$ in *eq. (7)*

	Else if layer l ==full connected layer then $a_j^l = sig\left(\sum_{i=1}^{n} x_i^{l-1} \times w_{ji}^l + b_j^l\right)$ in *eq. (3)* Else if layer l ==output layer then $p(y_s = j \mid \mathbf{a}_s; \boldsymbol{\theta}) = \frac{e^{\theta_j^{\mathrm{T}} \mathbf{a}_s}}{\sum_{j=1}^{K} e^{\theta_j^{\mathrm{T}} \mathbf{a}_s}}$ in *eq. (4)* End End Using the above steps, the segmentation map of the image $\mathbf{Y}_{CWNN}$ is obtained.
Superpixel Labeling:	Using the watershed method to obtain the superpixels of the image; Compute the histogram of the labels in each superpixel; Assign the superpixel to the label corresponding to the histogram peak, then a new segmentation map $\hat{\mathbf{Y}}_{CWNN}$ is obtained.
Fusion:	Using *eq.(8)* to obtain the segmentation map $\mathbf{Y}_{MRF}$; Compute the sketch map of the image; The region of edge and details is labeled by $\mathbf{Y}_{MRF}$, the rest region is labeled by $\hat{\mathbf{Y}}_{CWNN}$.

6. Experiments and analyses

In this section, the proposed approach is evaluated quantitatively and qualitatively on texture images and real SAR images. Firstly, we give some introduction about the experimental data. Then, our approach is tested on the texture images. Finally, the approach is further verified on the real SAR images. The main steps of our approach are shown in Table 1. Moreover, six methods are used for comparison. They include

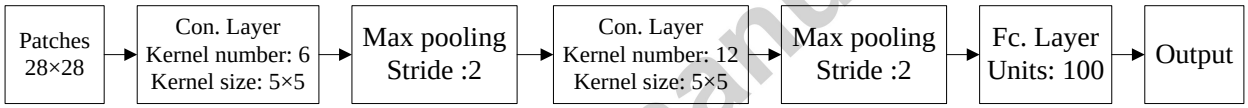
(1) MRF, it is a statistical model based approach and the MRF approach is introduced in section 3.

(2) CRF, it is also a statistical model based method and a discriminative model. It directly defines the posterior probability of the class label conditioned on the observed data as a Gibbs distribution. CRF mainly includes the unary and pairwise potentials. The unary potential is trained by the multiclass SVM and the pairwise potential is described as multilevel logistic model. MAP is used to obtain the labels of the image [19].

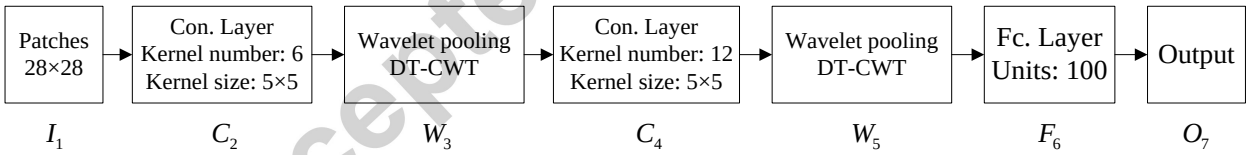
(3) GLCM, it is a kind of hand-engineered feature. The GLCM of the training patches are extracted, and then the features extracted from GLCM are used to train the softmax classifier. After the training, the testing patches are fed into the trained softmax classifier to obtain the labels of the image [7]. The size of the training and testing patches is 28×28.

(4) Gabor, Gabor filters are designed to be 5 scales and 8 directions. It means that the dimension of the feature vector for each image patch is 40. We extract the Gabor features from the training patches. Then, the extracted features are used to train the softmax classifier. The Gabor features extracted from all the patches of the image are fed into the trained softmax classifier to obtain the labels of the image [8].

(5) CNN, it is the original version of CNN. Its structure and parameters are given below [23]



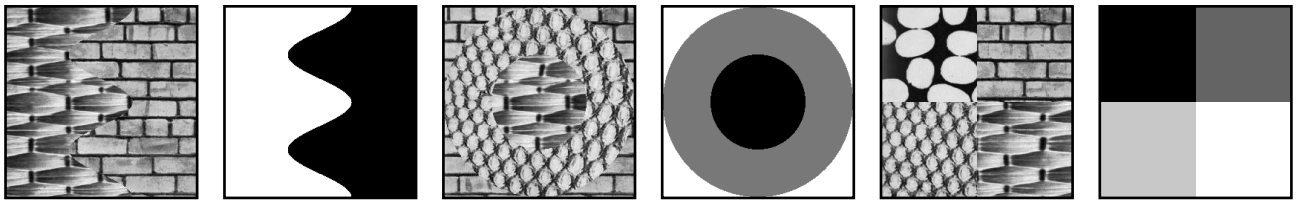
(6) CWNN, it is proposed in our paper. Its network architecture and parameters are given as follows



6.1. Test data sets

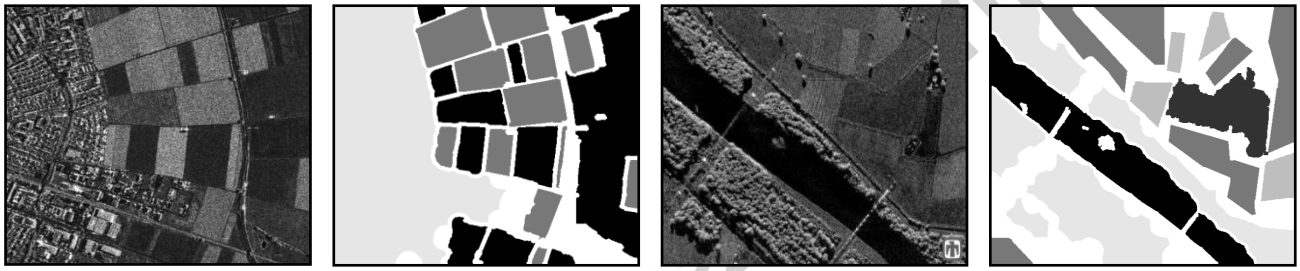
We test these approaches on two databases. One is the texture images including the two-class image, three-class image and four-class image. Each part of the image is from Brodatz database, which can be downloaded from http://www.cipr.rpi.edu/resource/stills/brodatz.html. These texture images are shown in Fig. 9. Another database is the real SAR images, including Noerdlinger Ries and Piperiver. They can be downloaded from www.dlr.de. and http://www.sandia.gov/radar/imagery/index.html., respectively. The corresponding dataset is shown

in Fig. 10. The white regions in Fig. 10(b) and (d) are the unlabeled regions and the rest are the labeled regions.



(a) (b) (c) (d) (e) (f)

Fig. 9. Texture images and their corresponding ground truth segmentations. (a) Two-class image; (b) Ground truth segmentation of the two-class image; (c) Three-class image; (d) Ground truth segmentation of the three-class image; (e) Four-class image; (f) Ground truth segmentation of the four-class image.



(a) (b) (c) (d)

Fig. 10. Noerdlinger Ries, Piperiver and their corresponding ground truth segmentations. (a) Noerdlinger Ries; (b) Ground truth segmentation of Noerdlinger Ries; (c) Piperiver; (d) Ground truth segmentation of Piperiver.

For the datasets used in our paper, we randomly select 40% labeled patches with the size of 28×28 as training samples. All the patches centered on the pixels with the size of 28×28 are used as the testing samples. The same training samples and the testing samples are used for the compared approaches. More information about the datasets will be introduced in Table 2 and Table 3, respectively.

Table 2

Prior information and number of samples in texture images.

Synthetic texture images				
Image	size	Class	Training samples	Testing samples
Two-Class	256×256	Class 1	12000	65536
		Class 2	12000	
Three-Class	256×256	Class 1	8000	65536
		Class 2	8000	
		Class 3	8000	
Four-Class	256×256	Class 1	6000	65536
		Class 2	6000	
		Class 3	6000	
		Class 4	6000	

Table 3

Prior information and number of samples in real SAR images.

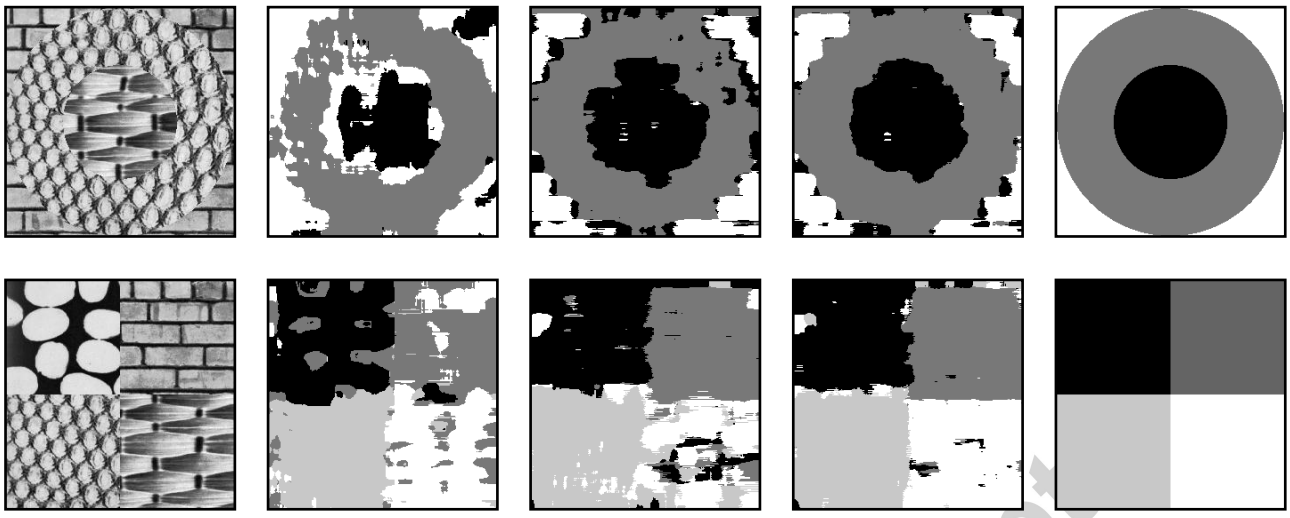
Real SAR images						
Image	size	Resolution	Band	Class	Training samples	Testing samples
Noerdlinger Ries	440×500	1	X	Class 1	20000	220000
				Class 2	20000	
				Class 3	20000	
Piperiver	432×600	1	Ku	Class 1	20000	259200
				Class 2	20000	
				Class 3	20000	
				Class 4	20000	
				Class 5	20000	

6.2. Segmentation results of the texture images

In order to verify the performance of CWNN for segmentation task, we test GLCM, CNN and CWNN on three texture images. The segmentation results of GLCM, CNN, and CWNN are given in Fig. 11(b), (c), and (d), respectively. Fig. 11(e) is the ground truth segmentation. We can see the segmentation results of CNNs (CNN and CWNN) are better than that of GLCM. That is because that GLCM is a hand-engineered feature and extracts the features at the single scale and the single

orientation. That is to say GLCM can not well represent the images with different scales and orientations. Unlike GLCM, CNNs automatically learn the features with stronger representation power. By comparing the results of CNN (shown in Fig. 11(c)) and CWNN (shown in Fig. 11(d)), we find that the segmentation results of CWNN are superior to the results of CNN. This shows that the wavelet pooling is effective and it improves the performance of the network. Such a pooling method will make the average of the low frequency subbands of DT-CWT to be used in replacing the conventional pooling. This strategy will not only suppress some noise but it will also keep the structures of the feature maps in lined with the specified rules. This new network architecture obviously improves the segmentation results. Moreover, from the segmentation results, we find that CNNs are not good at the edge localization. The reason is that the larger patch leads to the edges generalization.

We also evaluate our approach by utilizing two numerical indexes: 1) OA (overall accuracy) is the percentage of pixels that are correctly segmented; 2) Kappa coefficient is computed from the confusion matrix of the result and is used to measure the labeling consistency of the segmentation result. The numerical results are given in Table 4. The numbers in the brackets are the dimensions of the feature vectors. We can see that the accuracy of GLCM is lower than the accuracy of CNNs. The accuracy of CWNN is higher than that of the original CNN. In conclusions, the new network architecture CWNN is effective and improves the segmentation performance.



(a) (b) (c) (d) (e)

Fig. 11. Segmentation results of texture images. (a) Texture images; (b) Segmentation results by GLCM; (c) Segmentation results by CNN; (d) (d) Segmentation results by CWNN; (e) Ground truth segmentation.

Table 4

Accuracy (in %) and Kappa coefficients for synthetic texture images.

Method / Image	GLCM (12)		CNN (192)		CWNN (192)	
	OA	Kappa	OA	Kappa	OA	Kappa
2-Class	88.54	0.7706	93.97	0.8794	**96.18**	**0.9297**
3-Class	79.21	0.6479	83.11	0.7134	**88.66**	**0.8016**
4-Class	79.89	0.7319	81.28	0.7504	**92.36**	**0.8981**

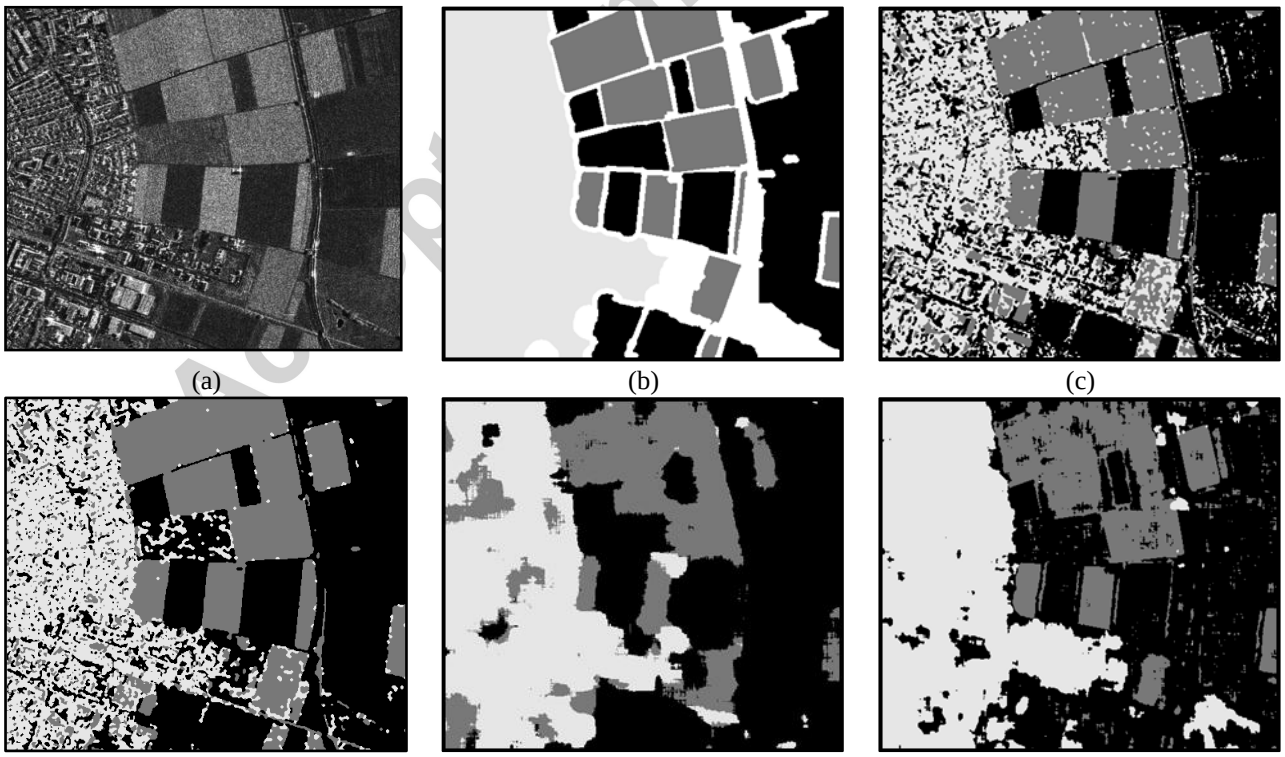
6.3. Segmentation results of the real SAR images

In this section, two real SAR images Noerdlinger Ries and Piperiver are used for a further analysis. The segmentation results are shown in Fig. 12 and Fig. 13. Fig. 12(a) and 13(a) are Noerdlinger Ries and Piperiver images, respectively. Fig. 12(b) and 13(b) are the corresponding ground truth segmentations. The segmentation results of MRF are shown in Fig. 12(c) and 13(c). It is observed that the edges and details are well preserved, but some over-segmentation occurs, especially in urban areas and forests. The reason is that the predefined spatial context model encounters some difficulties when capturing the complex textures in the SAR image. Fig. 12(d) and

13(d) give the segmentation results of the CRF model. CRF model is similar to MRF model. However, CRF model can capture the interactions in both of the observed data and the labels. Hence, the segmentation results of CRF are slightly better than that of MRF model. The segmentation results of GLCM are shown in Fig. 12(e) and 13(e). It is noted that the homogeneous regions (such as farmland and river) are segmented very well and the errors mainly occur in the heterogeneous regions (such as urban area and forest). The reason is that GLCM features are extracted in a single scale and orientation. Fig. 12(f) and 13(f) give the segmentation results of Gabor features. The Gabor features are extracted in five scales and eight orientations. Hence, the segmentation results of the urban area and forest are better than GLCM. It is noteworthy that Fig. 12 (g)-(h) and Fig. 13(g)-(h) are the segmentation results after the superpixel approach is performed. From the results of CNN (shown in Fig. 12(g) and 13(g)), we can see that the labeling consistency of the heterogeneous land covers is obviously better than that of the previous methods. In CWNN method (shown in Fig. 12(h) and 13(h)), the segmentation result is better than that of the original CNN. This shows that the performance of the network is improved by introducing the wavelet pooling. The new network architecture learns a better representation for the input image patch. From the results of our approach (shown in Fig. 12(i) and 13(i)), we note that various classes such as water, lands, urban areas and forests are identified clearly. The labeling consistency of the region is satisfied and the edges and details are well preserved at the same time. This is because that CWNN has strong representation power and improves the labeling consistency of the segmentation results. Moreover, two labeling strategies are used to refine the segmentation map of CWNN. In summary, CWNN is used to improve the labeling consistency of the region and the MRF is used to preserve the edges and the details.

The numerical results of Noerdlinger Ries and Piperiver are listed in Table 5 and Table 6, respectively. The last columns are the dimensions of the feature vectors. According to the network

architecture in Fig. 5, the dimension of the feature vector of CNN and CWNN are the same (192). For Gabor filter, we use five scales and eight orientations. Hence, the dimensional of the feature vector of Gabor is 40. In GLCM, we extract 12 textual features. Hence, the dimensional of the feature vector of GLCM is 12. It should be noted that the accuracies of our final results (shown in Fig. 12(i) and 13(i)) are not shown in the lists. That is because the accuracy of our approach and CWNN are the same in the labeled part. The difference between them is the unlabeled part. The quantitative evaluations in the unlabeled part are not given due to the lacking of the ground truth segmentation. In Table 5 and 6, the accuracy of the heterogeneous land covers by CNN and CWNN is dramatically higher than that of the approaches with hand-engineered features. Although the accuracies of some classes are lower than the other methods. The average accuracy and Kappa coefficients are the highest by our approach. This demonstrates that our approach has a good performance in most cases.



(a) (b) (c)

(d) (e) (f)

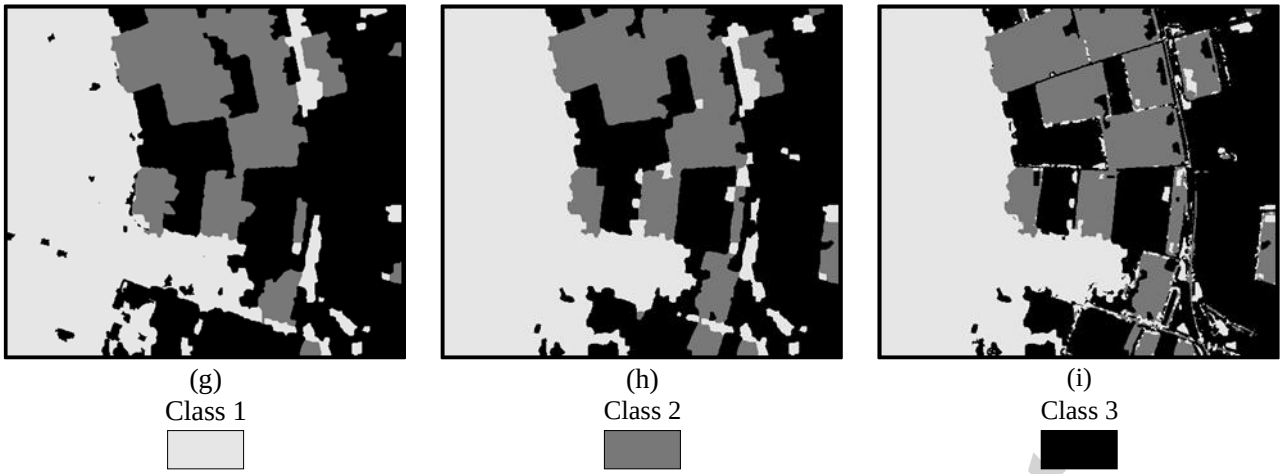


Fig. 12. Segmentation results of Noerdlinger Ries. (a) Noerdlinger Ries; (b)Ground truth segmentation; (c) Supervised MRF; (d) Supervised CRF; (e) GLCM; (f) Gabor; (g) CNN; (h) CWNN; (i) Our method.

Table 5

Accuracy (in %) and Kappa coefficients for Noerdlinger Ries.

Methods	Class1	Class2	Class3	Average	Kappa	Dimension
CWNN	**99.01**	95.47	90.54	**95.01**	**0.9351**	192
CNN	95.40	96.03	85.29	92.24	0.8954	192
Gabor	91.58	96.43	81.59	89.87	0.8573	40
GLCM	82.81	**98.13**	83.00	87.98	0.8167	12
CRF	67.03	94.99	**97.72**	86.58	0.7797	-
MRF	61.22	88.94	92.28	80.81	0.6674	-

(a) (b) (c)

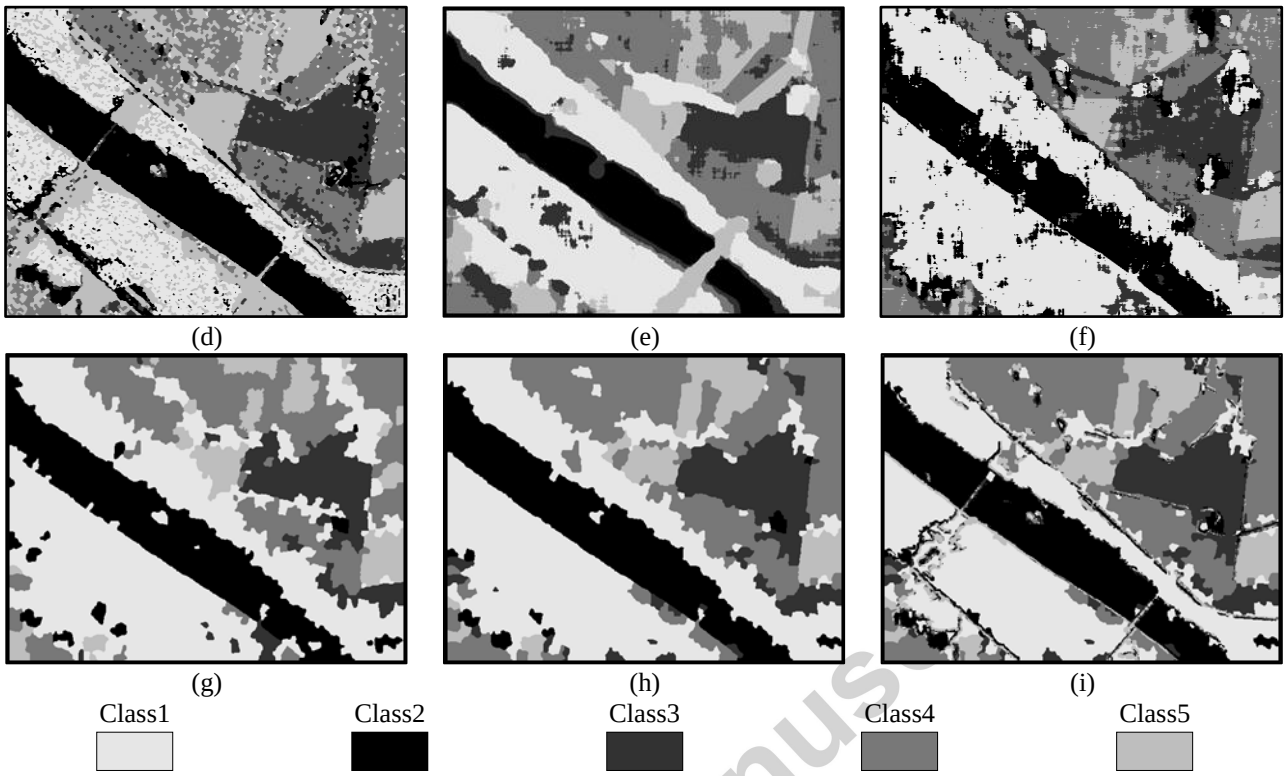


Fig. 13. Segmentation results of Piperiver. (a) Piperiver; (b) Ground truth segmentation; (c) Supervised MRF; (d) Supervised CRF; (e) GLCM; (f) Gabor; (g) CNN; (h) CWNN; (i) Our method.

Table 6

Accuracy (in %) and Kappa coefficients for Piperiver.

Methods	Class1	Class2	Class3	Class4	Class5	Average	Kappa	Dimension
CWNN	**94.43**	97.95	**96.40**	88.30	86.44	**92.70**	**0.9024**	192
CNN	94.33	96.39	90.29	86.71	79.37	89.42	0.8723	192
Gabor	89.75	86.94	82.94	87.51	88.63	87.16	0.8393	40
GLCM	87.40	85.33	87.44	83.36	**88.91**	86.49	0.8080	12
CRF	70.72	**98.81**	86.58	94.26	80.53	86.18	0.7502	-
MRF	66.96	98.85	79.09	86.64	79.88	82.28	0.7059	-

6.4. Parameter analysis

The size of the input patch is a critical parameter in our proposed approach. In this section, we analyze the effect of the input patch size on the segmentation results. We fix other parameters of the CWNN and the input patch size is a single variable. The input image patch size measures the local contextual information of the image captured by the network. For texture images and SAR images,

if the patch size is too small, it is not enough to capture the local contextual information of the image. Moreover, CWNN always produces the discriminative features to distinguish one class from another. If the patch size is too small, the input patches from different classes may be similar. It is difficult for CNN to learn the features to distinguish them. However, if the patch size is too big, it will increase the burden of the network and consume much more time for training and testing. It means that the segmentation result with neither too small nor too large input patch size will be satisfied. In order to illustrate the question clearly, we give some experiments to show the effect of the input patch size on the segmentation results of two-class texture image. The two-class texture image and its ground truth segmentation are shown in Fig. 11. The segmentation results with the input patches 16×16, 20×20, 24×24, 28×28 and 32×32 are shown in Fig. 14(a)-(e), respectively. We can see that the labeling consistency is improved by increasing the patch size. Moreover, the segmentation results with the input patches 16×16 and 20×20 are not very good. That is because that the local contextual information is not enough for the classification. The segmentation results are improved by increasing the patch size. The numerical indexes are shown in Table 7. We can see that after the patch size of 28×28, the accuracy increases slowly, but the larger patch will increase the burden of the network and consume much more time for training and testing. In order to balance the accuracy and the running time, we selected the size of the input patch to be 28×28.

(a) (b) (c) (d) (e)

Fig. 14. Segmetation results with different sizes of the input image patch. (a) Segmentation result with the input patch size 16×16; (b) Segmentation result with the input pacth size 20×20; (c) Segmentation result with the

input pacth size 24×24; (d) Segmentation result with the input pacth size 28×28; (e) Segmentation result with the input pacth size 32×32.

Table 7

OA (%) and Kappa coefficients for the two-class texture image with different sizes of the input image patch.

size	16×16		20×20		24×24		28×28		32×32	
Accuracy	OA	Kappa	OA	Kappa	OA	Kappa	OA	Kappa	OA	Kappa
	81.36	0.6276	86.25	0.7253	94.78	0.9056	96.18	0.9297	97.00	0.9399

6.5. Computational complexity of the proposed approach

For CWNN, it consists of convolutional layer, pooling layer and full connected layer. However, the pooling layer and full connected layer often take 5-10% of the computational time [49]. Most of the computational time is consumed by the convolutional layers. According to [49], the total time complexity of all convolutional layers was

$$O\left(\sum_{l=1}^{d} t_{l-1}\times v_l^2\times t_l\times m_l^2\right) \tag{12}$$

where l is the index of the convolutional layer, and d is the depth (the number of convolutional layers). t_l is the number of filters in the l th layer and t_{l-1} is also known as the number of input channels of the $(l-1)$ th layer. Moreover, v_l and m_l are the size of the filter and the output feature map, respectively.

In fact, our proposed approach consists of three parts, CWNN, the superpixel approach and the MRF approach. Hence, the computational complexity of our proposed approach is the sum of the computational complexity of CWNN, the MRF approach and the superpixel approach.

Although our proposed approach has the high computational complexity, it achieves the best results among these compared approaches. Therefore, our proposed approach deserves to be studied due to its promising results.

7. Conclusions

In this paper, we have proposed a novel SAR image segmentation method based on convolutional-wavelet neural network and Markov random field. In our proposed method, a wavelet constrained pooling layer has been proposed to replace the conventional pooling layer. The new network architecture produces a better representation of the input image patch. Hence, CWNN improves the performance of SAR image segmentation. Moreover, two labeling strategies are used to produce the final the segmentation results. It maintains that the labeling consistency is obtained and the edges and details are preserved at the same time. The experiments on texture images and real SAR images demonstrate that CWNN is effective for image segmentation task. In addition, the segmentation results on the real SAR images verify that our approach improves the labeling consistency and the preservation of the edges and the details.

In our future work, we will explore on other neural networks and their applications for SAR image segmentation. In addition, the characteristics of SAR image will be taken into account, and these characteristics will be plugged into the neural network for better segmentation results.

ACKNOWLEDGMENT

The work was carried out with the part-supports of the National Basic Research Program (973 Program) of China (No.2013CB329402), the National Natural Science Foundation of China (No.61573267, 61571342, 61572383 and 61601274), the Program for Cheung Kong Scholars and Innovative Research Team in University (No.IRT_15R53), the Fund for Foreign Scholars in University Research and Teaching Programs (the 111 Project) (No.B07048), the Major Research Plan of the National Natural Science Foundation of China (No.91438201 and 91438103), the Fundamental Research Funds for the Central Universities (No.JB140317).

REFERENCES

[1] R. H. Nobre, F. A. A. Rodrigues, R. C. P. Marques, J. S. Nobre, J. F. S. R. Neto, and F. N. S. Medeiros, SAR image segmentation with Renyi's entropy, IEEE Signal Processing Letters, 23(11) (2016) 1551-1555.

[2] F. Liu, Y. P. Duan, L. L. Li, L. C. Jiao, J. Wu, S. Y. Yang, X. R. Zhang, J. L. Yuan, SAR image segmentation based on hierarchical visual semantic and adaptive neighborhood multinomial latent model, IEEE Trans. Geosci. Remote Sens., 54(7) (2016) 4287-4301.

[3] L. C. Jiao, M. G. Gong, S. Wang, B. Hou, Z. Zheng, and Q. D. Wu, Natural and remote sensing image segmentation using memetic computing, IEEE Comp. Int. Mag. 5(2) (2010) 78-91.

[4] H. Yu, X. R. Zhang, S. Wang, and B. Hou, Context-based hierarchical unequal merging for SAR image segmentation, IEEE Trans. Geosci. Remote Sens., 51(2) (2013) 995-1009.

[5] L. C. Jiao, X. Tang, B. Hou, and S. Wang, SAR images retrieval based on semantic classification and region-based similarity measure for earth observation, IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., 8(8) (2015) 3876-3891.

[6] U. Kandaswamy, D. A. Adjeroh, and M. C. Lee, Efficient texture analysis of SAR imagery, IEEE Trans. Geosci. Remote Sens., 43(9) (2005) 2075-2083.

[7] R. M. Haralick, and K. Shanmugam, Texture features for image classification, IEEE Transactions on systems, man, and cybernetics, (6) (1973) 610-621.

[8] H. Yu, L. C. Jiao, and F. Liu. CRIM-FCHO: SAR image two-stage segmentation with multifeature ensemble, IEEE Trans. Geosci. Remote Sens., 54(4) (2016) 2400-2423.

[9] D. A. Clausi, Comparison and fusion of co-occurrence, Gabor and MRF texture features for classification of SAR sea ice imagery, Atmos. Ocean, 39(3) (2001) 183-194.

[10] F. A. A. Rodrigues, J. F. S. R. Neto, R. C. P. Marques, et al., SAR image segmentation using

the roughness information, IEEE Geosci. Remote. Sens. Lett., 13(2) (2016) 132-136.

[11] J. Gu, L. C. Jiao, S. Y. Yang, F. Liu, B. Hou, and Z. Q. Zhao, A multi-kernel joint sparse graph for SAR image segmentation, IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., 9(3) (2016) 1265-1284.

[12] O. Germain, and P. Réfrégier, Edge location in SAR images: Performance of likelihood ratio filter and accuracy improvement with an active contour approach, IEEE Trans. Image Process., 10(1) (2001) 72-78.

[13] R. Touzi, A. Lopes, and P. Bousquet, A statistical and geometrical edge detector for SAR images, IEEE Trans. Geosci. Remote Sens., 26(6) (1998) 764-773.

[14] B. Ogor, V. Haese-coat, and J. Ronsin, SAR image segmentation by mathematical morphology and texture analysis, In Proc. IEEE Int. Geosci. Remote Sens. Symp., Lincoln, 1, (1996) 717-719.

[15] C. Tison, J. M. Nicolas, F. Tupin, et al., A new statistical model for markovian classification of urban areas in high-resolution SAR images, IEEE Trans. Geosci. Remote Sens., 42(10) (2004) 2046-2057.

[16] P. Zhang, M. Li, Y. Wu, L. Gan, M. Liu, F. Wang, G. F. Liu, Unsupervised Multi-class segmentation of SAR images using fuzzy triplet Markov fields model, Pattern Recognition, 45(11) (2012) 4018-4033.

[17] Q. Y. Yu, and D. A. Clausi, IRGS: Image Segmentation Using Edge Penalties and Region Growing, IEEE Trans. Pattern Anal. Mach. Intell., 32(12) (2008) 2126-2139.

[18] A. Voisin, V. A. Krylov, G. Moser, S. B. Serpico and J. Zerubia, Classification of very high resolution SAR images of urban areas using copulas and texture in a hierarchical Markov random field model, IEEE Geosci. Remote. Sens. Lett., 10(1) (2013) 96-100.

[19] P. Zhang, M. Li, Y. Wu, and H. J. Li, Hierarchical conditional random fields model for

semisupervised SAR image segmentation, IEEE Trans. Geosci. Remote Sens., 53(9) (2015) 4933-4951.

[20] K. Kayabol, and J. Zerubia, Unsupervised Amplitude and Texture Classification of SAR Images with Multinomial Latent Model, IEEE Trans. Image Process., 22(2) (2013) 561-572.

[21] V. A. Krylov, G. Moser, S. B. Serpico, and J. Zerubia, On the method of logarithmic cumulants for parametric probability density function estimation, IEEE Trans. Image Process., 22(10) (2013) 3791-3806.

[22] A. C. Frery, H. J. Muller, C. C. F. Yanasse, et al., A model for extremely heterogeneous clutter, IEEE Trans. Geosci. Remote Sens., 35(3) (1997) 648-659.

[23] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE, 86(11) (1998) 2278-2324.

[24] 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.

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

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

[27] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, Face recognition: a convolutional neural-network approach, IEEE. Trans. Neural Networks, 8(1) (1997) 98-113.

[28] W. Ouyang, X. Wang, X, Zeng, et al., DeepID-Net: deformable deep convolutional neural networks for object detection, CVPR, (2015).

[29] X. Y. Zhang, J. H. Zou, K. M. He, and J. Sun, Accelerating very deep convolutional networks

for classification and detection, IEEE Trans. Pattern Anal. Mach. Intell., 38(10) (2016) 1943-1955.

[30] H. S. Li, R. Zhao, and X. G. Wang, Highly efficient forward and backward propagation of convolutional neural networks for pixelwise classification, Technical report, Electronic engineering at the Chinese University of Hong Kong, (2014).

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

[32] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, Semantic image segmentation with deep convolutional nets and fully connected CRFs, ICLR, (2015).

[33] P. H. O. Pinheiro, and R. Collobert. Recurrent convolutional neural networks for scene parsing, ICML, (2014).

[34] J. Long, E. Shelhamer, and T. Darrell, Fully convolutional networks for semantic segmentation, CVPR, (2015) 3431-3440.

[35] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, Image coding using wavelet transform, IEEE Trans. Image Process., 1(2) (1992) 205-220.

[36] S. G. Mallat, A theory for multiresolution signal decomposition: the wavelet representation, IEEE Trans. Pattern Anal. Mach. Intell., 11(7) (1989) 674-693.

[37] J. Wu, Fang Liu, L. C. Jiao, X. R. Zhang, H. X. Hao and S. Wang, Local maximal homogeneous region search for SAR speckle reduction with sketch-based geometrical kernel function, IEEE Trans. Geosci. Remote Sens., 52(9) (2014) 1-14.

[38] Y. Lecun, K. Kavukcuoglu, C. Farabet, Convolutional networks and applications in vision, ISCAS, (2010) 253-256..

[39] L. Zhang, W. P. Ma, D. Zhang, Stacked sparse autoencoder in PolSAR data classification using local spatial information, IEEE Geosci. Remote. Sens. Lett., 13(9) (2016) 1359-1363.

[40] L.C. Yang, B.Leon, B. O. Generieve, and M. Klaus-Robert, Efficient backprop, Neural networks: Tricks of the trade. Springer Berlin Heidelberg, (2012) 9-48.

[41] N. Kingsbury, Image processing with complex wavelets, Philosophical Transactions of the Royal society of London A: mathematical, physical and engineering sciences, 357(1760) (1999) 2543-2560.

[42] N. Kingsbury, The dual tree complex wavelet transform: a new efficient tool for image restoration and enhancement, Proceedings of EUSIPCO, (1998) 1-4.

[43] N. Kingsbury, Complex wavelets for shift invariant analysis and filtering of signals, Applied and computational harmonic analysis, 10(3) (2001) 234-253.

[44] I. W. Selesnick, R. G. Baraniuk, N. G. Kingsbury, The dual-tree complex wavelet transform-a coherent framework for multiscale signal and image processing, IEEE Signal Processing Magazine,22(6) (2005) 123-151.

[45] L. Vincent, and P. Soille, Watersheds in digital spaces: an efficient algorithm based on immersion simulations, IEEE Trans. Pattern Anal. Mach. Intell., 13(6) (1991) 583-598.

[46] F. Liu, L. P. Lin, L. C. Jiao, L. L. Li, S. Y. Yang, B. Hou, H. M. Ma, and J. H. Xu, Nonconvex compressed sensing by nature-inspired optimization algorithms, IEEE Trans. Cybern., 45(5) (2015) 1028-1039.

[47] C. Shi, F. Liu, L. L. Li, L. C. Jiao, Y. P. Duan, and S. Wang, Learning interpolation via regional map for pan-sharpening, IEEE Trans. Geosci. Remote Sens., 53(6) (2015) 3417-3431.

[48] F. Liu, J. F. Shi, L. C. Jiao, H. Y. Liu, S. Y. Yang, J. Wu, H. X. Hao, and J. L. Yuan, Hierarchical semantic model and scattering mechanism based PolSAR image classification, Pattern recognition, 59 (2016) 325-342.

[49] K. M. He, and J. Sun, Convolutional neural network at constrained time cost, CVPR, 2015.

Author biographies

Yiping Duan received the B. S. degree at the school of computer science and technology, Henan Normal University, Xinxiang, China, in 2010. She is currently working towards the Ph.D. degree at the School of Computer Science and technology, Xidian University, Xi'an, China. Her current research interests include semantic mining, machine learning and SAR image processing.

Fang Liu (SM'07) received the B.S. degree in computer science and technology from the Xi'an Jiaotong University, Xi'an, China, in 1984, and the M.S. degree in computer science and technology from the Xidian University, Xi'an, in 1995.

She is currently a Professor at Xidian University, Xi'an, China. She has authored or co-authored of five books and over 80 papers in journals and conferences. Her current research interests include image perception and pattern recognition, machine learning, evolutionary computation and data mining. She won the second prize of National Natural Science Award in 2013.

Licheng Jiao (SM'89) received the B.S. degree from Shanghai Jiaotong University, Shanghai, China, in 1982 and the M.S. and Ph.D. degrees from Xi'an Jiaotong University, Xi'an, China, in 1984 and 1990, respectively.

He is currently a Distinguished Professor with the School of Electronic Engineering, Xidian University, Xi'an. He has led approximately 40 important scientific research projects and has published over 10 monographs and 100 papers in international journals and conferences. He is the Author of three books Theory of Neural Network Systems (Xidian University Press, 1990), Theory and Application on Nonlinear Transformation Functions (Xidian University Press, 1992), and Applications and Implementations of Neural Networks (Xidian University Press, 1996). He has authored or coauthored of over 150 scientific papers. His current research interests include signal and image processing, natural computation, and intelligent information processing.

Prof. Jiao is a member of the IEEE Xi'an Section Executive Committee and the Chairman of the Awards and Recognition Committee, and an Executive Committee Member of the Chinese Association of Artificial Intelligence. He won the second prize of National Natural Science Award in 2013.

Peng Zhao received the B.S. degree at the School of Computer Science and technology, Northwestern Polytechnical University, Xi′an, China, in 2014. Now he is currently working towards

the M. S. degree at the School of Computer Science and technology of Xidian University. He current research interests include neural network, SAR image segmentation and dictionary learning.

Lu Zhang received the B.S. degree at the School of Electronic Engineering, Xidian University, Xi′an, China, in 2011. Now she is currently working towards the Ph.D. degree at the School of Electronic Engineering of Xidian University. Her current research interests include neural network, polarimetric SAR classification and multiobjective optimization.

Highlights

(1) SAR imaging system is usually an observation of the earths' surface. It means that rich structures exist in SAR images. Convolutional neural network (CNN) is good at learning features from raw data automatically, especially the structural features. Hence, CNN is used to improve the segmentation performance of SAR images.

(2) The conventional pooling (max pooling) in CNN do not take into the structure of the previous layer, blindly taking the max in the rectangular window as the output. In our approach, a wavelet constrained pooling layer is designed to replace the conventional pooling. The wavelet pooling extracts the features according to the specified rules.

(3) The wavelet pooling is plugged into the original CNN to generate the new network architecture named CWNN. The new network architecture can suppress the noise and is better at keeping the structures of the learned features. This contributes a more reliable segmentation.

(4) CWNN is not sufficient for the accurate segmentation. CWNN based on patches always produce the poor localizations of edges and small details. In order to remedy the problems, two labeling strategies are used to refine the segmentation results. Specifically, the labeling strategy based on the superpixel is used to force the local region into the same label. The labeling strategy based on the sketch map is used to locate the edges and details accurately.