EURASIP Journal on Image and Video Processing

RESEARCH

Open Access

Research on road extraction of remote sensing image based on convolutional neural network

Yuantao Jiang

Abstract

Road is an important kind of basic geographic information. Road information extraction plays an important role in traffic management, urban planning, automatic vehicle navigation, and emergency management. With the development of remote sensing technology, the quality of high-resolution satellite images is improved and more easily obtained, which makes it possible to use remote sensing images to locate roads accurately. Therefore, it is an urgent problem to extract road information from remote sensing images. To solve this problem, a road extraction method based on convolutional neural network is proposed in this paper. Firstly, convolutional neural network is used to classify the high-resolution remote sensing images into two classes, which can distinguish the road from the non-road and extract the road information initially. Secondly, the convolutional neural network is optimized and improved from the training algorithm. Finally, because of the influence of natural scene factors such as house and tree shadow, the non-road noise still exists in the road results extracted by the optimized convolutional neural network method. Therefore, this paper uses wavelet packet method to filter these non-road noises, so as to accurately present the road information in remote sensing images. The simulation results show that the road information of remote sensing image can be preliminarily distinguished by convolutional neural network; the road information can be distinguished effectively by optimizing convolutional neural network; and the wavelet packet method can effectively remove noise interference. Therefore, the proposed road extraction method based on convolutional neural network has good road information extraction effect.

Keywords: Road extraction, Remote sensing image, Convolutional neural network, Training algorithm, Wavelet packet method

1 Introduction

Geographic information system (GIS) has been widely used in many fields. At present, the acquisition and real-time update of GIS data is the most important bottleneck problem restricting the application of GIS [1, 2]. Traditional GIS data come directly from manual work site survey and mapping. With the development of sensing technology and remote sensing technology, it is possible to use remote sensing satellite images to locate and identify the ground objects in GIS. This provides new ideas and means for the dynamic acquisition and real-time updating of GIS data and then the development of digital measurement and mapping, mobile target recognition, and traffic planning. In recent years, the spatial resolution of satellite sensors has been improved continuously. For example, the spatial resolution of IKONOS-2 panchromatic band and near-infrared band images has reached 1 m and 4 m, and the spatial resolution of QuickBird panchromatic band and multi-spectral band images has reached 0.61 m and 2.44 m. High-resolution remote sensing satellite images can observe the detailed changes of the earth's surface on a small space scale. Using high-resolution remote sensing satellite images to accurately locate and identify the ground objects has become a hot research topic in the field of space science.

Road is a kind of important basic geographic information, which undoubtedly occupies a pivotal position in

Correspondence: ytjiang@shmtu.edu.cn
School of Economics and Management, Shanghai Maritime University, Shanghai, China

urban land use and economic activities. Highly accurate and timely updating of road network information plays a very important role in traffic management, urban planning, automatic vehicle navigation, and emergency management [3]. Using the computer to extract road information from remote sensing images can not only update the road network information in time to achieve dynamic data acquisition, but also can be used for reference for the extraction of other linear features [4]. The research of road information extraction by computer can be traced back to the 1970s, and some achievements have been made in the field of low-resolution remote sensing images [5]. Some famous software companies and research institutions, more representatives of RSI and NASA, developed Feature Extension module and ICREST software, respectively, and achieved good results in low-resolution image applications. However, in high-resolution satellite images, due to the complex topological structure of the road and abundant scene information, it is often affected by many kinds of noise, such as road vehicles, white lines, trees, and buildings, so the effect is not satisfactory. Road extraction from high-resolution remote sensing image is still a difficult problem [6, 7].

To extract information from remote sensing images, the object and image features must be distinctly extracted and translated into semantic description. Image features are formed by the physical and geometric features of the scene which make the local areas of the image change significantly. The road's image features mainly include geometric, radiative, topological, and contextual features. The road on the high-resolution image shows a narrow and continuous region with slow change of gray scale and geometric features. The main characteristics of the ideal road are as follows [8–10]:

1. Geometric features: width consistency, large aspect ratio, small curvature. The overall width of the road is relatively uniform. The ratio of length to width of a road in an image can be understood as greater than a certain constant. For the road, the safety and feasibility of vehicle turning are considered in the design, and the bending rate is generally small and changes slowly.
2. Radiation characteristics: mainly for the gray characteristics, that is, the road gray level is relatively consistent and changes slowly, and the road surface and background has a certain contrast.
3. Topological characteristics: roads are connected all the time, forming interconnected road networks, and there will be no abrupt interruption in the middle.
4. Contextual features: contextual features refer to the image features associated with the road. The buildings and road trees along the road have been relatively single, not covering the road surface, and can be used to judge whether the urban road or the rural expressway.

The reduction of visual difference between high spatial resolution remote sensing image and natural image provides the condition for the application of convolutional neural network in the field of remote sensing target recognition. The great success of convolutional neural network in natural image target recognition also promotes the application of convolutional neural network in remote sensing target recognition. However, remote sensing images are more complex and changeable than natural images. The large image size, image distortion, object occlusion, shadow coverage, illumination changes, and texture heterogeneity and other factors bring great challenges to the practical application of convolutional neural network in high spatial resolution remote sensing image target recognition. How to apply convolutional neural network to target recognition of high spatial resolution remote sensing image is of great theoretical significance.

Based on the good performance of convolutional neural network, this paper applies convolutional neural network to road extraction of remote sensing image, puts forward a road extraction method of remote sensing image based on convolutional neural network, and effectively extract road information from remote sensing images. The main contributions of this paper are as follows:

1. Convolutional neural network is applied to road information extraction from remote sensing images.
2. The convolutional neural network is used to classify the high-resolution remote sensing images into two classifications, to distinguish the road from the non-road and to preliminarily extract the road information.
3. The convolutional neural network is optimized and improved from training algorithm.
4. Wavelet packet method is used to filter out non-road noise caused by natural scenes such as houses and tree shadows.

2 Related work

2.1 Convolutional neural network

Krizhevsky et al. (2012) [11] proposed the Alex Net model. For the first time, convolutional neural network was applied to large-scale natural image classification, and won the champion of image classification group in the 2012 Large Scale Visual Recognition Challenge (ILSVRC). The error rate of 16.4% is about 10 percentage points lower than that of the traditional algorithm. Consequently, the convolution neural network has set off a research upsurge in the field of computer visionsuch as natural image classification, target recognition, image segmentation . Since then, VGGNet (Simonyan and Zisserman, 2014) [12], GoogLeNet (Szegedy et al., 2016) [13] and other convolution neuralnetwork models have been put forward one after another, constantly refreshing the ILSVRC competition-records. Girshick and Donahue (2014) [14] first applied convolutional neural networks to target recognition inlarge scale natural images. The proposed R-CNN model obtained an average accuracy of about 20%higher than the traditional algorithm. R-CNN model lays the foundation of target recognition usingconvolutional neural network. Subsequently, the improved R-CNN model, such as Fast R-CNN (Girshick etal, 2015) [15], Faster R-CNN (Ren et al, 2016) [16], has been proposed one after another, which improves theaccuracy and speed of target recognition significantly.

2.1.1 The structure of convolutional neural network

Convolutional neural network (CNN) is a kind of feed-forward neural network, which is an efficient identification method developed in recent years [23]. It was first proposed by Hubel and Wiesel on the basis of the study of visual cortex electrophysiology of cats. Convolutional neural network is composed of input layer, convolution layer, pool layer, and full connection layer. It is different from general neural network. On the one hand, a sole meridian element in convolutional neural network is often connected only with some adjacent neurons, that is, the structure of local connection. On the other hand, some weights are shared between neurons, which means that convolutional neural networks have fewer weights, and the complexity of the network is lower, which is closer to biological neural networks.

Its general structure can be categorized as follows:

a. Input layer

Input layer is the input of the whole network. Input of computer vision problems is usually a picture, an image is a matrix of pixels. The height and width of the input layer correspond to the height and width of the matrix. The depth of the input layer is the number of channels of input image. Starting from the input layer, the input image is sampled by different convolution layer and pool layer and transformed into a layer of different matrix feature maps. Finally, the high-dimensional features of the image are extracted, which are used for classification and subsequent processing.

b. Convolution layer

As the name implies, the most important layer of convolutional neural networks is the convolution layer. Convolution layer filters the sub-nodes of each layer to extract abstract features. Each node in the convolution layer only uses a small part of the output of the previous layer to connect and calculate. The size of the filter is the size of CNN through the convolution layer for feature extraction.

The filter is a matrix whose length and width are specified by hand, usually 3 × 3 or 5 × 5, and the individual network can reach 11 × 11. Through convolution calculation, the output layer of the current layer is obtained, that is, the convolution layer propagates forward. The filter transforms the nodes in the receptive field of the current layer into a unit node matrix of the next layer only passing through the network. The unit node matrix is a node matrix with the same length and width and the same depth as the filter.

The storage and computation process of convolution layers involves some parameters. These parameters need to be defined when constructing the network. Data is defined as a blod and is a quaternion (a, b, c, d), representing the number of channels, height/row, and width/column, respectively. The structure of the filter can also be defined by such a four tuple.

In order to reduce unnecessary details, the convolution process can be carried out across pixels. Filter parameter step size strides can be specified. The step length refers to the size of the filter sliding at intervals on the convolution layer. Each time the filter moves one step, it calculates a unit node matrix, moves the filter from the upper left corner of the input layer to the lower right corner, and all the calculated unit node matrices are combined into a new matrix. The output layer of the convolution is obtained.

The formula for convolutional neural network is as follows:

$$a_{i,j} = f\left(\sum_{m=0}^{2}\sum_{n=0}^{2} W_{m,n}X_{m+i,n+j} + W_b\right) \tag{1}$$

where $X_{i,\ j}$represents the pixel value of line i in column j of the image, $W_{m,\ n}$ represents the weight of column n in row m, and W_b represents the bias of the filter. Similarly, $a_{i,\ j}$ represents the value of column j of the i row of the feature map; the activation function is represented by f.

c. Pooling layer

After calculating the weights and offsets of the convolution layer, the eigenvalues of the image are obtained, which is the basis of classification. But the data after the multi-layer convolution layer is huge. It is necessary to reduce the size of the matrix by using the pool layer, reduce the parameters of the final full-connected layer, and achieve the purpose of reducing the amount of calculation. On the other hand, the pooling layer can also prevent data from overfitting.

The calculation of the pooled layer is a kind of data resampling. Similar to the convolution layer, the forward propagation is accomplished by a matrix shift similar to a filter. The difference is that the pool is not filtered, and the maximum or average value is calculated. The most common pooling methods are max pooling [24], mean pooling [25], and random pooling. The maximum pooling layer is called max pooling layer, and calculated by the average is average pooling layer. The random pooling is selected randomly according to the probability matrix.

d. Fully connected layer

Fully connected layer (FC) is a traditional multi-layer perceptron network in the last part of the whole model. Unlike the local connectivity of convolution layer neurons, each layer in FC uses global information of the previous layer, that is, each node's calculation is connected with the weights of all nodes in the previous layer. After convolution and pooling operations, the output of the network is the high-dimensional feature information of the image, so the full connection layer is used to classify the features. A convolutional neural network usually has two or three full-connection layers and a classifier. After the extraction of convolution layer information, the fusion, and classification of full-connection layer information, the output is the probability that the image corresponds to a certain label.

2.1.2 Characteristics of convolutional neural networks

Convolutional neural network is a kind of multilayer artificial neural network. It simulates the visual system of biology and realizes the invariance of image feature displacement, scaling, and distortion by three techniques: sparse connection, weight sharing, and down-sampling. Sparse connection can extract local, primary visual features; weight sharing ensures that the network has fewer parameters; de-sampling reduces the resolution of features and achieves invariance to displacement, scaling, and rotation.

a. Sparse connection

Sparse connection is the neural units in different layers that are locally connected, that is, the neural units in each layer are only connected to a part of the nerve unit in the previous layer. Each neural unit only responds to the region within the receptive field and does not care at all about the area outside the receptive field. This local connection mode ensures that the learning convolution check input spatial local pattern has the strongest response.

With the deepening of network layers, the area covered by high-level neural cells is becoming larger and larger, thus abstracting the global feature information of the image. Compared with the method of extracting global features from each unit, this method greatly reduces the network parameters and computational complexity and enables the model to deal with more complex target recognition tasks. Moreover, with the deepening of the network, the underlying area of the high-level neurons is getting larger and larger. At the same time, the simple features extracted from the underlying layer are combined to form more abstract and complex features in the high-level.

b. Weight sharing

Weight sharing means that each convolution kernel in the convolution layer is convoluted with the whole feature map to generate a new feature map. Each cell in the feature graph shares the same weight and bias parameters.

Suppose the M-1 level i characteristic graph has 5 units, that is, there are 5 input units in the m layer. The size of the sparse connection in the m layer is 3. The convolution kernel $k = (k_1, k_2, k_3)$ with three weight parameters is used to convolution the input feature map, and the output feature map of m layer is obtained from three hidden layer units. Although each cell in the m layer is connected to only three cells in the m-1 layer, a total of nine connections are made, and the same color connections share the same convolution kernel. Weight sharing makes the features extracted by convolutional neural network independent of the position of input space, that is, the features of the convolutional neural network have translation invariance. In addition, as with sparse connections, weight sharing reduces the number of learning parameters in convolutional neural network model again and improves the efficiency of the algorithm.

c. Descending sampling

In the convolutional neural network model, the pool layer usually follows the convolution layer, and the output characteristic map of the convolution layer is sampled down. De-sampling greatly reduces the number of neural cells in the intermediate layer between the input layer and the output layer. If the sampling size is $m \times n$, the number of cells in the input layer of size $M \times N$ is reduced to $M \times N/(m \times n)$after sampling, thus reducing the complexity of model calculation. Furthermore, the de-sampling operation makes the convolutional neural network model have memory function, makes the extracted features have certain distortion and deformation invariance, and enhances the generalization ability of features.

2.2 Wavelet packet transform method

The following is the introduction of the principle of the wavelet packet algorithm. According to the principle of the wavelet packet, the wavelet packet decomposition is shown in Fig. 1.

In Fig. 1, A indicates the low-frequency part, D indicates the high-frequency part, and the end number indicates the number of layers of wavelet packet decomposition. The relationship between decomposition is as follows:

$$S = AAA3 + DAA3 + ADA3 + DDA3 + AAD3 + DAD3 + ADD3 + DDD3 \tag{2}$$

In multi-resolution analysis, $L^2(R) = \underset{j\in Z}{\oplus} W_j$. It is shown that multi-resolution analysis decomposes the Hilbert space $L^2(R)$ into the orthogonal sum of all subspaces $W_j (j \in Z)$ according to different factors j, where W_j is the closure of the wavelet function $\psi(t)$ (wavelet subspace). Now, we need to further subdivide the wavelet subspace W_j according to the binary fraction to improve the frequency resolution. A common approach is to unify the scale subspace V_j and the wavelet subspace W_j with a new subspace U_j^n, if

$$\begin{cases} U_j^0 = V_j \\ U_j^1 = W_j \end{cases}, j\in Z \tag{3}$$

Then, the orthogonal decomposition $V_{j+1} = V_j \oplus W_j$ of Hilbert space can be unified by the decomposition of U_j^n

.

$$U_{j+1}^0 = U_j^0 \oplus U_j^1, j\in Z \tag{4}$$

Define subspace U_j^n is the closure space of function $u_n(t)$, and U_j^{2n} is the closure space of function $u_{2n}(t)$, and let $u_n(t)$ satisfy the following two-scale equation:

$$\begin{cases} u_{2n}(t) = \sqrt{2}\sum_{k\in Z} h(k) u_n(2t-k) \\ u_{2n-1}(t) = \sqrt{2}\sum_{k\in Z} h(k) u_n(2t-k) \end{cases} \tag{5}$$

where $g(k) = (-1)^k h(1-k)$, that is, the two coefficients are orthogonal to each other. When $n = 0$, given directly from the above two formulas:

$$\begin{cases} u_0(t) = \sqrt{2}\sum_{k\in Z} h(k) u_0(2t-k) \\ u_1(t) = \sqrt{2}\sum_{k\in Z} h(k) u_0(2t-k) \end{cases} \tag{6}$$

In the multi-resolution analysis, $\phi(t)$ and $\psi(t)$ satisfy the dual scale equation.

$$\begin{cases} \phi(t) = \sum_{k\in Z} h_k \phi(2t-k), \{h_k\}_{k\in Z} \in l^2 \\ \psi(t) = \sum_{k\in Z} g_k \phi(2t-k), \{g_k\}_{k\in Z} \in l^2 \end{cases} \tag{7}$$

$u_0(t)$ and $u_1(t)$ are reduced to scale function $\phi(t)$ and wavelet basis function $\psi(t)$ respectively. This equivalent expression is extended to the case of $n \in Z_+$(non-negative integer)

$$U_{j+1}^n = U_j^n \oplus U_j^{2n+1}, j\in Z; n\in Z_+ \tag{8}$$

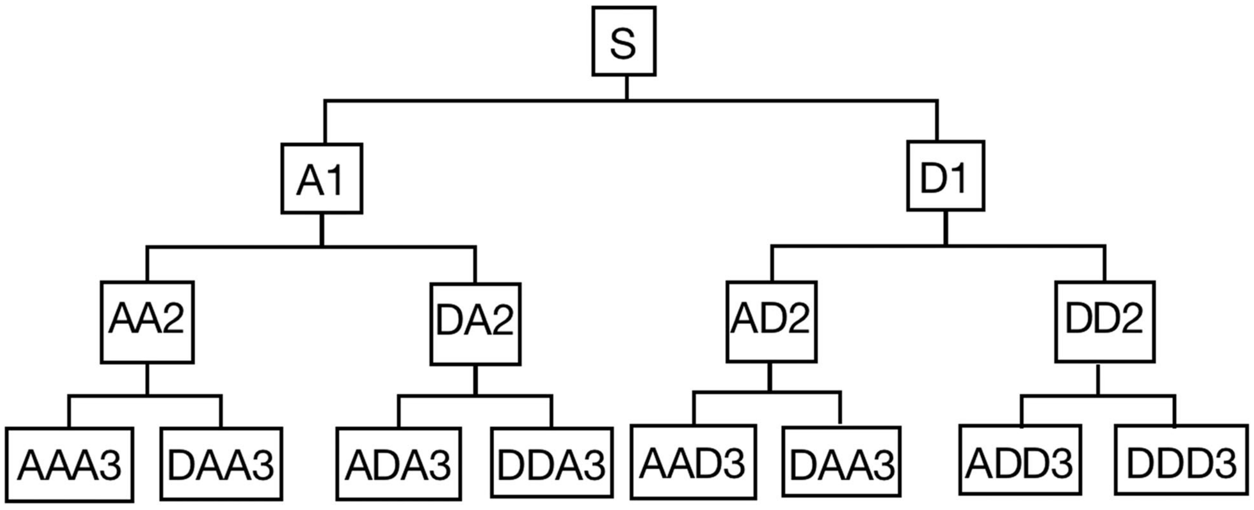


Fig. 1 Wavelet packet decomposition graph

3 Proposed method

3.1 Training process of convolutional neural network

In this paper, the road extraction of remote sensing image is studied. Firstly, the convolutional neural network is used to classify the high-resolution remote sensing image, distinguish the road from the non-road, and extract the road information initially. Secondly, the convolutional neural network is optimized and improved from the training algorithm. In this section, the training process of convolutional neural network is improved to realize road classification in remote sensing images.

Gradient descent algorithm is simple, easy to converge, but easy to fall into the local optimal solution, and the gradient descent near the saddle point is slow, affecting the training of network model. The saddle point can be avoided in the training process of Newton algorithm, but the method needs to compute Heisen matrix to ensure its non-negative positive definite and a large amount of storage space and to ensure the existence of second-order derivatives of the objective function; otherwise, it is difficult to ensure the convergence of the algorithm [26–28]. In order to avoid the difficulties caused by the direct use of the Newton algorithm, the improved BFGS quasi-Newton algorithm [29] is used to train convolutional neural networks.

In the Newton algorithm training process, the CNN model parameters are updated to:

$$W_{i+1} = W_i - \frac{\nabla j_w}{\nabla^2 j_w} \tag{9}$$

The initial quasi-Newton algorithm equation is:

$$B_{k+1} S_k = y_k \tag{10}$$

where $S_k = W_{k+1} - W_k$, $y_k = \nabla j_{w+1} - \nabla j_w$. In order to achieve better target optimization effect, the improved BFGS is used in this paper.

$$H_{k+1} = H_k + \frac{y_k y_k^t}{y_k^t s_k} - \frac{H_k s_k s_k^t H_k}{s_k^t H_k s_k} \tag{11}$$

However, the Heisen matrix obtained by the recursion formula may be a singular matrix, and the inverse matrix calculation has a certain degree of complexity, so the recursion formula in this paper is expressed as follows:

$$B_{k+1}^{-1} = \left(B_0 - \frac{s_k y_k^T}{s_k^T y_k} \right) B_k^{-1} \left(B_0 - \frac{y_k s_k^T}{s_k^T y_k} \right) + \frac{s_k s_k^T}{s_k^T y_k} \tag{12}$$

Based on the above theory and formula, the training process of the improved BFGS algorithm is as follows:

1. Select iteration times epochs, epoch = 0, initialization parameter w and offset b.
2. If epoch < epochs, iteration stops.
3. Calculate the value of $b_k d_k +\quad j_w = 0$ and get a new optimization direction $d_k = -B_k^{-1} g_k$.
4. Search along the d_k direction to satisfy $w_{k+1} = w_k + a_k d_k$, it is the minimum point in this direction, $a_k > 0$.
5. Let the epoch iteration add 1 and go to (2).

3.2 Wavelet packet denoising

In the last section, the convolutional neural network method is used to extract roads from remote sensing images. However, due to the influence of natural scenes such as houses and tree shadows, there will be non-road noise in the extraction results. Therefore, this paper uses the wavelet packet method to filter these irrelevant non-road noises, so as to accurately put forward the road information in remote sensing images.

Wavelet analysis can decompose time-frequency effectively, but the resolution of the high-frequency part after processing is very poor, and the image information of the low-frequency part is incomplete [30]. The wavelet packet transform divides the signal frequency into two parts: low-frequency A1 and high-frequency D1. The lost information in low-frequency A1 is obtained by high frequency. In the next layer, A1 is decomposed into two parts: low-frequency A2 and high-frequency D2. The lost information in low-frequency A2 is obtained by high-frequency D2. Therefore, for deeper decomposition, wavelet transform processing will undoubtedly lose some details. The case of wavelet packet transform is different; it can provide a more accurate signal analysis; and it will not only decompose the frequency at different levels, but also subdivide the high-frequency part. Therefore, not only the low-frequency part is decomposed, but also the high-frequency part is decomposed, and the wavelet transformation cannot be achieved, only the wavelet packet method can deal with this.

3.2.1 Improved wavelet packet transform method

Let $U_j^0 = V_j$, $U_j^1 = W_j$; W_j is a closure generated by $\psi(t)$ of wavelet function, then the orthogonal decomposition $V_{j+1} = V_j \oplus W_j$ of Hilbert space can be decomposed into $U_{j+1}^0 = U_j^0 \oplus U_j^1, j \in Z$ by U_j^n. The definition of subspace U_j^n is the closure space generated by the function $u_n(t)$, U_j^{2n} is the closure space of function $u_{2n}(t)$, and make $u_n(t)$ satisfy the following two scale equations:

$$\begin{cases} u_{2n}(t) = \sqrt{2} \sum\limits_{k \in Z} h(k) u_n(2t - k) \\ u_{2n+1}(t) = \sqrt{2} \sum\limits_{k \in Z} h(k) u_n(2t - k) \end{cases} \tag{13}$$

where $g(k) = (-1)^k h(1 - k)$, that is, the two coefficient has orthogonal relations. Wavelet packets are defined as

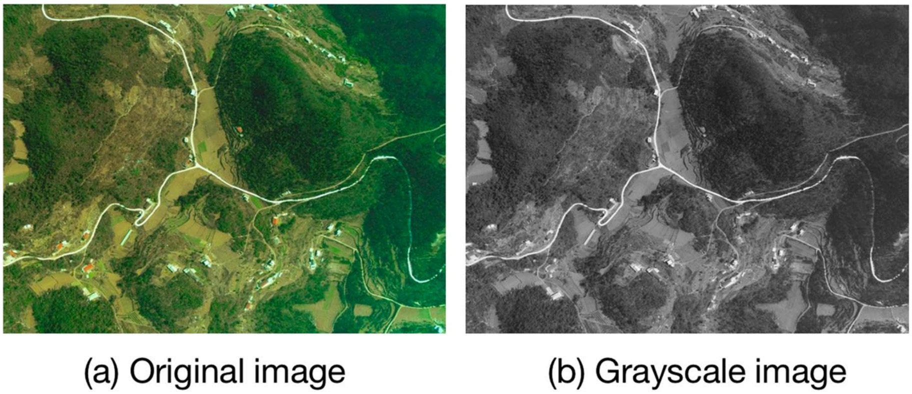


Fig. 2 Grayscale processing. **a** Original image. **b** Grayscale image

follows: the constructed sequence $\{u_n(t)\}(n \in Z_+)$ is called the orthogonal wavelet packet determined by the basis function $u_0(t) = \phi(t)$, the scaling function $\phi(t)$ is uniquely determined by the low pass filter$\{h(t)\}$. Let $g_j^n(t) \in U_j^n$, $g_j^n(t) = \sum_l d_l^{j,n} u_n(2^j t - l)$, the wavelet packet decomposition algorithm is as follows:

$$\begin{cases} d_l^{j,2n} = \sum_k a_{k-2l} d_l^{j+1,n} \\ d_l^{j,2n+1} = \sum_k b_{k-2l} d_l^{j+1,n} \end{cases} \tag{14}$$

$d_l^{j,2n}$ and $d_l^{j,2n+1}$ are calculated from $d_l^{j+1,n}$. The formula of wavelet packet reconstruction is as follows: $d_l^{j+1,n} = \sum_k [h_{l-2k} d_k^{j+1,2n} + g_{l-2k} d_k^{j+1,2n+1}]$. That is to calculate $\{d_l^{j+1,n}\}$ through $\{d_l^{j,2n}\}$ and $\{d_l^{j,2n+1}\}$, where $\{a_k\}$ and $\{b_k\}$are decomposed filters, $\{h_k\}$ and $\{g_k\}$ are reconfigurable filters, also called orthogonal mirror filter banks.

3.2.2 Fusion algorithm of wavelet packet transform

The frequency domain characteristic of the wavelet is that the image is decomposed into different frequency signals and multi-resolution layers. It fully reflects the change of the local area of the original image, thus providing favorable conditions for the image data. In the original image, the region with obvious change is the region with change of gray value, which shows the change of the absolute value of coefficients in the change of wavelet. So the key problem is how to get the characteristics of the image according to the wavelet packet coefficients. At present, there are two fusion methods of wavelet transform, one is single pixel fusion, the other is based on the local window activation measure to take the large criterion of fusion [31].

The specific fusion method in this paper is as follows: considering that the low-frequency component of the J level (assuming the decomposition level is J) after wavelet packet decomposition, which contains most of the information of the image, has a certain impact on the quality of image fusion, the fusion method based on local window energy is adopted [32]. Set C_F, C_X, and C_Y

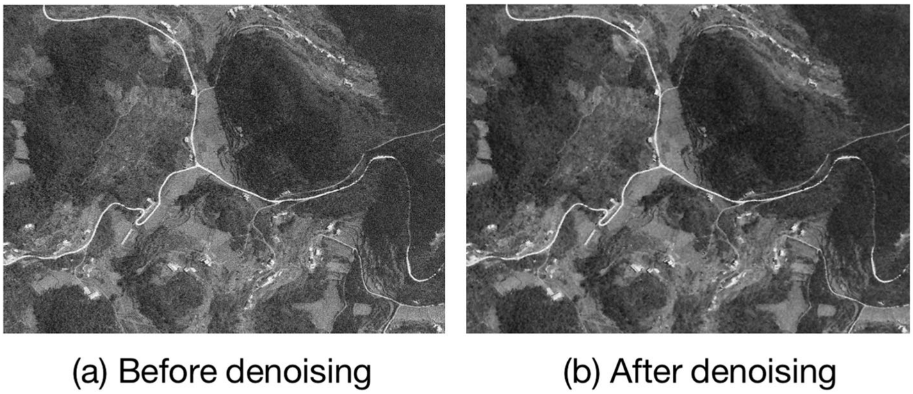
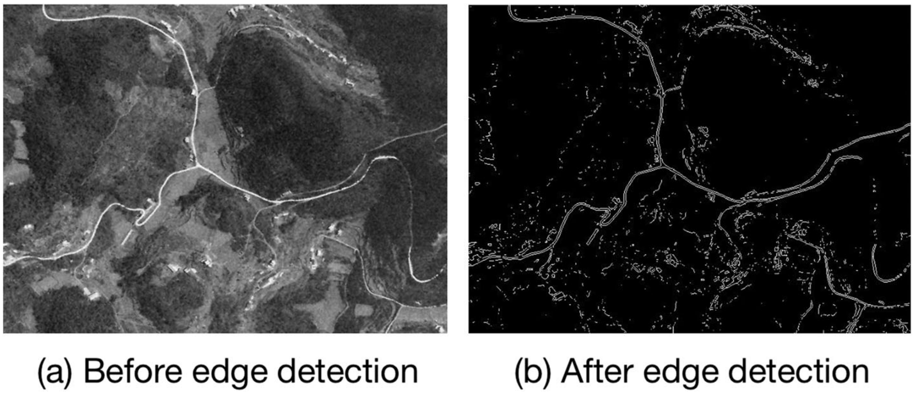


Fig. 3 De-noising processing. **a** Before de-noising. **b** After de-noising



(a) Before edge detection (b) After edge detection

Fig. 4 Edge detection. **a** Before edge detection. **b** After edge detection

to denote scale coefficients of fused image *F*, image *X*, and image *Y*respectively. m_X and m_Y are mathematical expectations of C_X and C_Y respectively. $N_k(u, v, n)$ represents a neighborhood of $n \times n$ size centered on point (u, v), in general, n is odd. The lower corner k is 1,2, which represents the scale coefficients of *X* and *Y*images respectively. $E_k(u, v, n)$ represents the energy of neighborhood $N_k(u, v, n)$,

$$E_k(u, v, n) = \left(\sum_i \sum_j (N(i, j) - m_k)^2 \right)^{\frac{1}{2}} / (n \times n) \tag{15}$$

The scale coefficient of the fused image can be calculated using the following expressions:

$$C_F(u, v) = W_1(u, v) \times C_X(u, v) + W_2(u, v) \times C_Y(u, v) \tag{16}$$

Where W_k is a weighting factor, the expression is: $W_k = E_k(u, v, n) / \sum_{m=1}^{2} E_m(u, v, n), k = 1, 2$.

Wavelet packet decomposition in high frequency region includes all details of image information [33]. However, the large absolute value coefficients correspond to the large-scale contrast in the image, such as image edge features, and the human eye is very sensitive to these features; so, its principle is to use the maximum selection least squares method. The mathematical expressions of this idea are as follows:

$$D_F(u, v) = \begin{cases} D_X(u, v), if \max\limits_{(u', v') \in N} \left(\left| D_X\left(u', v'\right) \right| \right) >= \max\limits_{(u', v') \in N} \left(\left| D_Y\left(u', v'\right) \right| \right) \\ D_Y(u, v), if \max\limits_{(u', v') \in N} \left(\left| D_X\left(u', v'\right) \right| \right) >= \max\limits_{(u', v') \in N} \left(\left| D_Y\left(u', v'\right) \right| \right) \end{cases} \tag{17}$$

(u, v) represents the spatial location of small coefficients, N represents the square window of a $n \times n$ centered on (u, v), (u', v') is an arbitrary point in the window N. In this paper, a remote sensing image method based on wavelet packet transform is proposed, which avoids the shortcoming that wavelet transform only decomposes the low-frequency components further, but does not consider the high-frequency components. The innovation is that the low-frequency part of the fusion algorithm is based on local window energy, while the high-frequency part is based on the maximum selection principle to calculate the wavelet packet coefficients. The fusion algorithm can effectively fuse the high-resolution image with the low-resolution image, keep the spectral characteristics of the original image as much as possible, and improve the spatial resolution of the fusion image.

4 Experimental results and discussions

It is difficult to extract road from remote sensing image. Before extracting road, it is necessary to preprocess image. Firstly, the image is grayed, as shown in Fig. 2. As can be seen from Fig. 2, the original image does not change the contour of the image after graying, but the image data changes from three dimensions to two dimensions.

Table 1 Road extraction index average evaluation data table

Method	Missing error (%)	Redundancy error (%)	Accuracy (%)
Deconvolution of convolutional neural network without wavelet packet method	8.34	0.87	87.25
Convolutional neural network and wavelet packet denoising	5.12	0.35	93.61

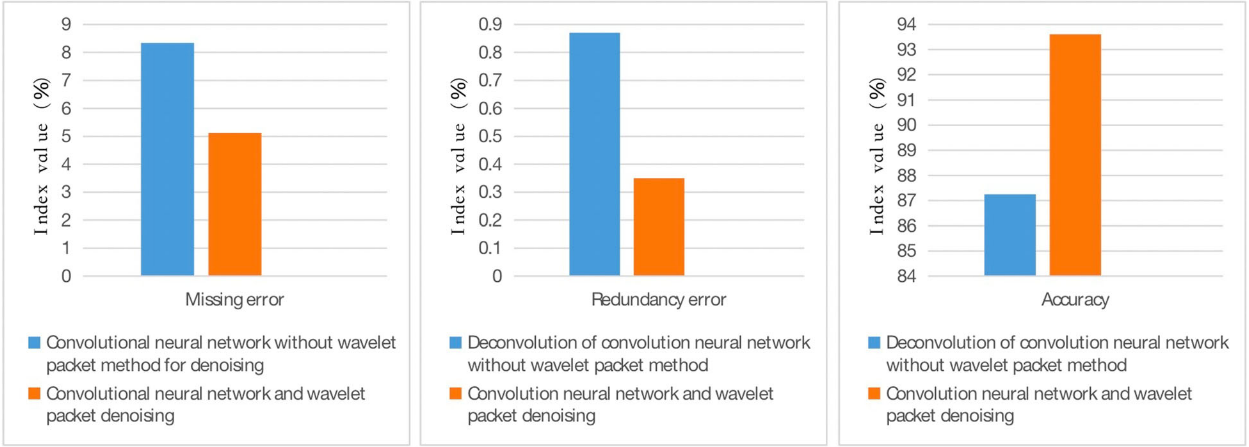


Fig. 5 Road extraction evaluation comparison histogram

After gray processing, in order to reduce the impact of noise, the image is denoised in this paper. The processing results are shown in Fig. 3. As can be seen from Fig. 3, after denoising, the image blurring is reduced and the image is clearer, which is helpful for subsequent image processing.

After denoising, this paper uses edge detection technology to process the image. The processing results are shown in Fig. 4. As can be seen from Fig. 4, besides the road edge detection curve, there are obvious non-road noises caused by natural scene factors such as house and tree shadows. Therefore, it is necessary to filter the non-road noise caused by natural scene factors such as house and tree shadows by using wavelet packet algorithm after the convolutional neural network algorithm is used to binary classify the image.

In order to better evaluate the extraction accuracy of the road extraction method designed in this paper, this paper further adopts the quantitative analysis method to evaluate the road extraction results, using three indexes of accuracy, omission error, and redundancy error to evaluate the road extraction results. The formulas for the three indicators are as follows:

Missing error = missing linear target length/total linear target length (18)

Redundant error = redundant linear target length/total linear target length (19)

Accuracy = correctly extracted linear target length/total linear target length (20)

In Formula (18), the missing linear target length is the length of the road as noise removal, and the redundant linear target length in Formula (19) is the length of the road as noise extraction. In this paper, 100 images are selected for processing, and the three performance indicators are used to evaluate the three indicators. The average evaluation data of the three indicators are shown in Table 1, and the data comparison histogram is shown in Fig. 5.

Combining Table 1 and Fig. 5, it can be seen clearly that only convolutional neural network can extract road information smoothly, but because of non-road noise caused by natural scene factors such as house and tree shadow, the values of the three indexes are relatively large. After re-processing with wavelet packet, it can be seen that the values of the three indexes decrease obviously, among which the omission error decreases by 3.22%, the redundancy error decreases by 0.52%, and the accuracy increases by 6.36%. The improvement of the values of the three indicators shows that the non-road noise caused by the shadows of houses and trees can be effectively reduced by the wavelet packet algorithm after the convolutional neural network is used to extract road information.

In order to verify the superiority of the proposed method, two other methods are compared with the proposed method. The two methods are Directional Morphological Enhancement and Segmentation (DMES) road detection method and adaptive Unsupervised

Table 2 Three methods evaluation datasheet

Method	Missing error (%)	Redundancy error (%)	Accuracy (%)
DMES	7.71	0.81	88.91
AUA	7.16	0.75	90.15
Method of this paper	5.12	0.35	93.61

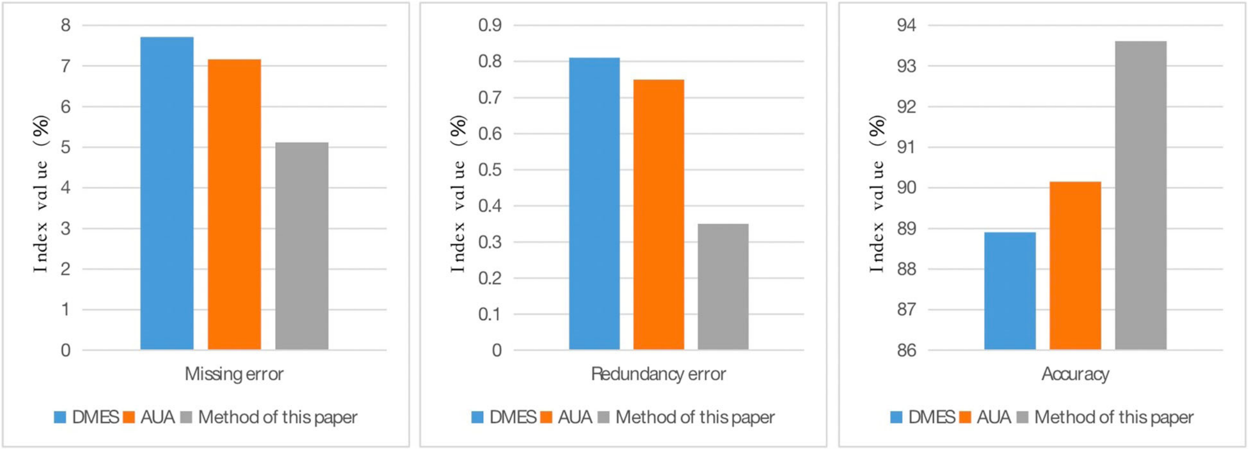


Fig. 6 Three methods for evaluating and comparing histogram

Approach (UA) road detection method. The three method compares the data as shown in Table 2, and the data contrast histogram is shown in Fig. 6.

Combining Table 2 and Fig. 6, it is obvious that the three methods are the worst in DMES, the second in AUA, and the best in this paper. Among them, the performance of DMES and AUA is not much different, AUA is 0.55% lower than DMES in missing error, AUA is 0.06% lower than DMES in redundancy error, and AUA is 1.24% higher than DMES in accuracy. The method proposed in this paper is better than DMES and AUA, and the missing error of this method is 2.04% lower than AUA, 2.59% lower than DMES, 0.4 lower than AUA, and 0.46% lower than DMES in redundancy error. In terms of accuracy, this method is 3.46 lower than AUA and 4.7 lower than DMES. Through numerical comparison, we can find that the proposed method has a great improvement in the performance of the three indicators, which shows that the proposed method has better road information extraction performance than DMES and AUA.

The simulation results show that the proposed road extraction method based on convolutional neural network has good performance on the basis of its effectiveness.

5 Conclusions

With the development of science and technology, remote sensing image is more and more easy to obtain, and the extraction of road information is conducive to traffic management, urban planning, automatic vehicle navigation, and emergency handling. In order to obtain better road information of remote sensing image, a road extraction method based on convolutional neural network is proposed in this paper. After optimizing the training algorithm of convolutional neural network, this paper uses wavelet packet algorithm to remove the non-road information interference caused by natural scene factors such as house and tree shadow and puts forward the road information in remote sensing image accurately. Through simulation experiments, this method is compared with DMES and AUA detection methods. The accuracy of this method is 4.7% and 3.46% higher than that of DMES and AUA, respectively. The simulation results show that the convolutional neural network can effectively classify the road and extract the road information, but there still exists the influence of non-road noise caused by the natural scene factors, such as houses, trees, and shadows. In this paper, the wavelet packet algorithm can effectively remove this kind of interference. In order to illustrate the superiority of this method, this paper uses the directional shape enhancement and damage separation road detection method and adaptive unsupervised road detection method to compare with this method, and it can be seen from the chart that this method has better road information extraction performance.

Abbreviations

CNN: Convolutional neural network; DMES: Directional Morphological Enhancement and Segmentation; FC: Fully connected layers; GIS: Geographic information system; ILSVRC: In the 2012 Large Scale Visual Recognition Challenge

Acknowledgements

The authors thank the editor and anonymous reviewers for their helpful comments and valuable suggestions.

About the Authors

Yuantao Jiang was born in Shandong, Taian, P.R. China, in 1975. He received the Doctor's degree from Huazhong University of Science and Technology, P.R. China. Now, he works in School of Economics and Management, Shanghai Maritime University. His research interest include business intelligence and innovation, information system and smart shipping.

Funding

This work was supported by Key Soft Science Project from the Shanghai Committee of Science and Technology, China (No.18692106700).

Availability of data and materials

We can provide the data.

Author's contributions

The author YJ wrote the first version of the paper, did all experiments of the paper, and revised the paper in different versions. The author read and approved the final manuscript.

Competing interests

The author declares that he has no competing interests

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 23 October 2018 Accepted: 15 January 2019
Published online: 01 February 2019

References

1. S. Yassemi, S. Dragićević, M. Schmidt, Design and implementation of an integrated GIS-based cellular automata model to characterize forest fire behaviour [J]. Ecol. Model. **210**(1), 71–84 (2017)
2. X. Liu, X. Wang, G. Wright, et al., A state-of-the-art review on the integration of building information modeling (BIM) and geographic information system (GIS) [J]. Int. J. Geo-Inf. **6**(2), 53 (2017)
3. J. Liu, Q. Qin, J. Li, et al., Rural road extraction from high-resolution remote sensing images based on geometric feature inference [J]. Int. J. Geo-Inf. **6**(10), 314 (2017)
4. M. Arifin, *Application of geographic information system for the road information (case study : Surabaya city) [J]. undergraduate theses* (2010)
5. Y. Wang, Y. Wang, Y. Da, et al., *An object-oriented method for road damage detection from high resolution remote sensing images [J]* (2011), pp. 1–5
6. X. Duan, Y. Zhang, Z. Guo, Feature extension of cluster analysis based on microblog [J]. Comput. Eng. Appl **53**(13), 90–95 (2017)
7. M.A. Huifang, X. Zeng, L.I. Xiaohong, et al., Short text feature extension method of improved frequent term set [J]. Comput. Eng. **42**(10), 213–218 (2016)
8. H. Pazhoumand-dar, M. Yaghoobi, A new approach in road sign recognition based on fast fractal coding. Neural Comput. & Applic. **22**(3–4), 615–625 (2013)
9. R.J. Wang, K. Wang, F. Zhang, et al., Emission characteristics of vehicles from national roads and provincial roads in China [J]. Environ. Sci. **38**(9), 3553 (2017)
10. K. Noh, J. Lee, J. Kim, Older drivers' characteristics and optimal number of guide names on road signs [J]. Transp. Res. Rec. J. Transp. Res. Board **2227**(−1), 71–78 (2008)
11. Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems. 1097–1105 (2012).
12. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. Computer Science, 2014.
13. Szegedy C, Vanhoucke V, Ioffe S, et al. [IEEE 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) - Las Vegas, NV, USA (2016.6.27-2016.6.30)] 2016 IEEE Conference on Computer. Vision and Pattern Recognition (CVPR) - Rethinking the Inception Architecture for Computer Vision[J]. 2016:2818–2826.
14. Girshick R, Donahue J, Darrell T, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[C]// 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society, 2014.
15. Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]// International Conference on Neural Information Processing Systems. 2015.
16. He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[C]// IEEE Conference on Computer Vision & Pattern Recognition. 2016.
17. S. Khawaldeh, U. Pervaiz, A. Rafiq, et al., Noninvasive grading of glioma tumor using magnetic resonance imaging with convolutional neural networks [J]. Appl. Sci. **8**(1), 27 (2017)
18. D. Shuang, A.R. Bugos, P.M. Hill, Design and evaluation of an Ethernet-based residential network [J]. IEEE J. Sel. Areas Commun. **14**(6), 1138–1150 (2002)
19. P. Ballester, R.M. Araujo, *On the performance of GoogLeNet and AlexNet applied to sketches[C]// Thirtieth AAAI Conference on Artificial Intelligence* (AAAI Press, 2016), pp. 1124–1128
20. N. Baharin, T.A.R.T. Abdullah, Challenges of PHEV penetration to the residential network in Malaysia [J]. Procedia Technology **11**(1), 359–365 (2013)
21. N. Zhang, J. Donahue, R. Girshick, et al., *Part-Based R-CNNs for fine-grained category detection [J]*, vol 8689 (2014), pp. 834–849
22. Z.Q. Zhao, H. Bian, D. Hu, et al., *Pedestrian detection based on Fast R-CNN and batch normalization [J]* (2017), pp. 735–746
23. L. Xu, J.S.J. Ren, C. Liu, et al., Deep convolutional neural network for image deconvolution [J]. Adv. Neural Inf. Proces. Syst. **2**, 1790–1798 (2014)
24. A. Giusti, C.C. Dan, J. Masci, et al., *Fast image scanning with deep max-pooling convolutional neural networks [J]* (2013), pp. 4034–4038
25. Y. Ma, H. Chen, G. Liu, General mean pooling strategy for color image quality assessment [J]. Laser Optoelectron. Prog. **55**(2), 021007 (2018)
26. X. Shao, X. He, *Statistical error analysis of the inverse compositional Gauss-Newton algorithm in digital image correlation [J]* (2017), pp. 75–76
27. P. Birtea, C. Dan, Newton algorithm on constraint manifolds and the 5-electron Thomson problem [J]. J. Optim. Theory Appl. **173**(2), 1–21 (2017)
28. C. Yin, S. Dadras, X. Huang, et al., Energy-saving control strategy for lighting system based on multivariate extremum seeking with Newton algorithm [J]. Energy Convers. Manag. **142**, 504–522 (2017)
29. B. Hao, L.S. Zan, BFGS quasi-Newton location algorithm using TDOAs and GROAs [J]. J. Syst. Eng. Electron. **24**(3), 341–348 (2013)
30. A.M. Babker, The methodology of wavelet analysis as a tool for cytology preparations image processing [J]. Cukurova Med J **41**(3), 453–463 (2016)
31. X. Xue, F. Xiang, H.A. Wang, Fusion method of panchromatic and multi-spectral remote sensing images based on wavelet transform [J]. J. Comput. Theor. Nanosci. **13**(2), 1479–1485 (2016)
32. X. Xiong, L. Ye, R. Yang, Distribution power system multi-objective optimization based on wind power wavelet packet decomposition and storage system fuzzy control [J]. Autom. Electr. Power Syst. **39**(15), 68–74 (2015)
33. L.U. Yun, X.U. Jun, Wind power hybrid energy storage capacity configuration based on wavelet packet decomposition [J]. Power Syst. Prot. Control **44**, 149–154 (2016)