



A novel improved deep convolutional neural network model for medical image fusion

Kai-jian Xia^{1,2} · Hong-sheng Yin¹ · Jiang-qiang Wang²

Received: 29 January 2018 / Accepted: 7 February 2018 / Published online: 22 February 2018
© Springer Science+Business Media, LLC, part of Springer Nature 2018

Abstract

This paper proposed a novel fusion scheme for multi-modal medical images that utilizes both the features of the multi-scale transformation and deep convolutional neural network. Firstly, the source images are decomposed by the Gauss-Laplace filter and Gaussian filter into several sub-images in the first layer of network. Then, HeK-based method is used to initialize the convolution kernel of the rest layers, construct the basic unit, and use the back propagation algorithm to train the basic unit; Train multiple basic units that are sacked with the thought of SAE to get the deep stacked neural network; the proposed network is adopted to decompose the input images to obtain their own high frequency and low frequency images, and combine the our fusion rule to fuse the two high frequency and low frequency images, and put them back to the last layer of the network to get the final fusion images. The performance of our proposed fusion method is evaluated by conducting several experiments on the different medical image datasets. Experimental results demonstrate that our proposed method does not only produce better results by successfully fusing the different images, but also ensures an improvement in the various quantitative parameters as compared to other existing methods. In addition, the speed of our improved CNN method is much faster than that of comparison algorithms which have good fusion quality.

Keywords Image fusion · Multi-scale transformation · Deep learning · Fusion rules · Auto encoder · CNN modal

1 Introduction

Medical imaging has been playing a very important role in the field of medical diagnosis since many years, which is a major source for the doctors to diagnose the diseases. Whatsoever the medical imaging has its own kinds of Imaging techniques like X-ray, computed tomography (CT), magnetic resonance imaging (MRI). However, the characteristics and results of each of these medical imaging techniques are unique. For instance, CT can provide images as dense like structure with which the physiological changes could not be detected whereas in MRI images even the soft pathological tissues can be visualized better. As a

result the anatomical and functional medical images are needed to be combined for better visualization and for accurate diagnosis [1–3].

Three different levels of image fusion are pixel level, feature level and decision level. The lowest level of fusion is pixel level in which fusion process is carried out on pixels. Images are segmented into regions and features like pixel intensities, edges or texture are used for fusion in feature level fusion. Assimilating information at a higher level of abstraction is seen in decision level fusion [4, 5]. Pixel-level fusion can retain as much raw detail information as possible, and provide the subtle information that other fusion methods cannot provide. The current image fusion research mainly focuses on the pixel level, where multi-scale transform fusion is the most active research field. From the early Discrete Wavelet Transform(DWT) to a series of improvements on the wavelet such as double-tree complex wavelet transform, up to the recent Non-subsampled Contourlet Transform(NSCT) [6–8] and Non-Subsampled Shearlet Transform (NSST), etc., the common thread of these methods is the Multi-scale

✉ Kai-jian Xia
xiakajian@163.com

¹ School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, Jiangsu, China

² Soochow University Changshu Hospital (Changshu No. 1 People's Hospital), Changshu 215500, Jiangsu, China

decomposition for source image, which can firstly obtain high-frequency and low-frequency components, and then respectively high and low frequency coefficients based on different fusion rules were combined, and finally the inverse transform of the combined high and low frequency components is achieved to obtain the final fusion results. In other words, these multi-modal images are decomposed on different frequency bands, then data compression and data decompositions are performed respectively [9]. Overall, the multi-scale transformation fusion mainly focuses on the following factors: (1) In order to prevent the loss of information, the original downsampling strategy is improved to non-subsampling. (2) In order to obtain the good local information such as the edge contour, the filtering direction is increased from the early three directions to the infinite directions. These improvements solve the ringing effect and shift-variant of wavelet transform, which causes the performance for image fusion has been significantly improved [10–15].

However, it has limited number of directions and large computational complexity, whenever it is used in image fusion scheme. To represent more edges, efficiently, Labate et al. introduced a new multiscale geometric analysis tool called shearlet which has all properties like other tools as multiscale, localization, anisotropy and directionality [16], but still it is not able to overcome the problem of shift invariance. Later, Easley et al. [17] proposed nonsampled shearlet transform (NSST) that is realized by non-sampled Laplacian pyramid (NSLP) and several shearing filters. The NSST provides the variable directional selectivity and shift invariance. In the recent years, various medical image fusion algorithms based on all these transformation techniques have been reported. In many cases, the acquisition of prior knowledge is difficult, and the essence of this problem is that the fixed model is difficult to adapt the fusion requirements of different fusion image. In view of the deep Learning has successfully broken the constraints in many fields. The literature [8] uses the deep support value to learn the network fusion for remote sensing images and achieves good results. However, the number of high-frequency images depends on the number of network layers, In addition, research also combine wavelet transform with low-frequency Auto Encoder, but it did not break the shortcoming of wavelet transform itself. Therefore, there is still some exploratory work for deep learning in medicine image fusion.

The study of image fusion has lasted for more than 30 years, during which hundreds of related scientific papers have been published. In recent years, deep learning (DL) has gained many breakthroughs in various computer vision and image processing problems. In the field of image fusion, the study based on deep learning has also become an active topic in the last three years. A variety of DL-

based image fusion methods have been proposed for digital photography (e.g., multi-focus image fusion, multi-exposure image fusion), multi-modality imaging (e.g., medical image fusion). Currently, the deep learning models that are widely recognized mainly include Auto encoder (AE), Deep Belief Network (DBN) and Convolutional Neural Network (CNN). AE is an unsupervised learning network, which encodes and decodes the signal within a certain error. Multiple AEs are stacked to make up a Stacked Auto Encoder (SAE). This kind of deep network makes it possible to learn the hierarchical expression of the input signal and realizes the compression and dimensionality reduction of the signal. DBN is a probabilistic generating model, consisting of multiple Restricted Boltzmann Machines (RBMs) that has been successfully used in speech recognition and computer vision. However, while DBN processing the input data, it does not consider the spatial structure of two-dimensional signal, which is not conducive to the processing of images and video signals. In CNN, convolution is used instead of the full connection of traditional neural network, combined with the sub-sampling of time or space, significantly reducing the number of free parameters in the network, thereby reducing the training complexity, gaining an outstanding effect for image recognition. Among the three methods, CNN extracts and abstracts the features on the images, through a series of convolution and downsampling, which is consistent with the idea of multi-scale decomposition of images by convolution and downsampling. The difference lies in that the filter requires defining manually in the multi-scale decomposition. Therefore, the bank of filters gained from deep convolutional neural network learning can solve series of problems resulting from relative fixity of the bank of filters in the multi-scale decomposition.

However, the following problems still exist for CNN: (1) For CNN, the eventually output feature through convolution and downsampling has lower dimensional degree, so using the model directly in image decomposing will lose information, leading to poor integration; (2) For the traditional CNN, the extracted features are not the features of high or low frequency for the image, which are not conducive to the specific formulation of fusion rules; (3) On the one hand, deep CNN is difficult to train; on the other hand, the learning ability of shallow CNN cannot guarantee the accuracy of image decomposing (4) Stacking CNNs through operations similar to stacking SAE, can resolve the problem referred in (3), but a small number of stacked units can not guarantee the accuracy, while a large number of units can lead to Network convergence difficulties; (5) If every type image is trained to get a DSCNN alone, the generalization ability of the network can be improved, but the reconstruction of fused images of high and low frequency is difficult to realize. Therefore, in this paper, the

down sampling layer of CNN is removed; a network is constructed only by using convolution layer based on Auto Encoder (AE), to realize that the output image is the same as the input image in the network, resolving the problem of losing information. Gauss Laplacian filter and Gaussian filter are used as the initial convolution kernels of the first layer network for the high-frequency sub-net and the low-frequency subnet respectively, after passing through the high-frequency sub-network and low-frequency sub-network, the high frequency image and low frequency image of the input image can be obtained; Deep Stacked Convolutional Neural Network (DSCNN), which takes CNN as its basic stacked unit, is used to improve the network learning ability; End-to-End training can improve the accuracy, stability and convergence of the network; Using different types of images to train the same DSCNN, can not only ensure the generalization ability of the network, but also facilitate to obtain the final fusion result.

The rest of the paper is organized as follows. Section 2 illustrates the related methodologies which are used to present the proposed fusion method. Section 3 presents the implementation of the proposed image fusion method that is based on the CNN. In Sect. 4, various experimental results are discussed and compared with the existing fusion methods on the basis of different performance measures. Section 5 depicts the final conclusions.

2 Related work

The following section of this article provides a brief introduction to multi-scale transformation and deep neural network before describing our architecture in more detail.

2.1 Deep stacked neural network

2.1.1 SAE model

AE is constituted by the input layer, the hidden layer and the output layer, and multiple AEs are stacked into a SAE, whose basic structure is shown in Fig. 1. Input $x \in R^{m_0}$ is encoded through the first layer, and the first-order feature of the data can be obtained, represented as $h_1 \in R^{m_1}$, and then input the first-order feature representation to the second layer automatic encoder to obtain the second-order feature, represented as $h_2 \in R^{m_2}$. Repeat the operation, and the representation of n th order feature $h_n \in R^{m_n}$ can be got. Different levels of feature after encoding constitute the hierarchical description of the original data. While training SAE, it only needs to be trained layer by layer in the way where it is trained by the single automatic encoder, and fine-tune the whole network with the error between the last layer output and expectation.

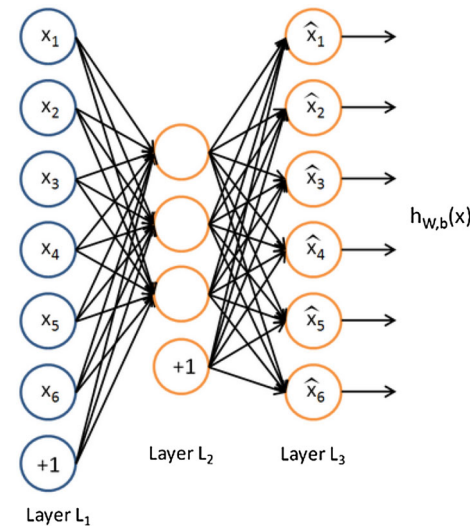


Fig. 1 SAE structure

2.1.2 CNN model

CNN consists of input layer, hidden layer and output layer. The hidden layer is mostly constituted by convolution layers and down-sampling layers alternately, where the convolution layer is used to extract the features of the image and the down-sampling layer is used to simplify the information. A simple structure of CNN is shown as Fig. 2, where an image is processed by three operations: convolution kernel, activation function and bias so as to generate the 3 features mapping images in C1 layer; through down-sampling, each features mapping image can get its feature mapping in S2 layer; S2 layer feature mapping is operated by convolution and downsampling, the feature mapping of C3 layer and S4 layer can be get; finally, in the last layer, all pixels are rasterized into a column of vector, which is input to the traditional Neural Network (NN) to get the output result. The network usually adopts the back propagation algorithm to train, including the forward propagation phase and the error back propagation phase. In the forward propagation, the samples are input from the input layer, passing through network layers one by one; then, the output is compared with the expected output to obtain the error, and then, the error is transmitted into the back-

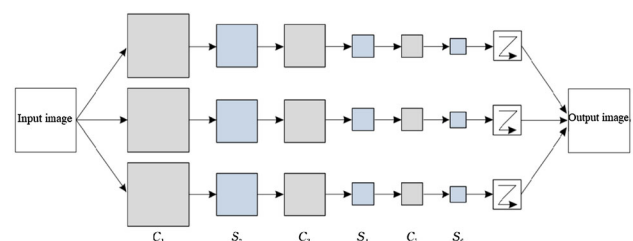


Fig. 2 Diagram of CNN structure

propagation phase. In the back-propagation phase, the error is passed forward from the output layer to the input layer one layer by one layer. For each layer, a current error should be obtained, which can be used as a basis for modifying the weight of each unit. Repeat the operation until the network reaches a certain precision, or the prescribed training time is expired.

2.2 Multi-scale transform

As described above, there are many methods for multi-scale image transformation, where DTCWT, NSCT and NSST are widely more adorable. The following section will provides a brief introduction to multi-scale transformation.

2.2.1 DTCWT

DTCWT is a commonly used signal processing method. It calculates the complex transform of a signal using two separate DWT decompositions (tree a and tree b). If the filters used in one are specifically designed different from those in the other, it is possible for one DWT to produce the real coefficients and the other the imaginary. DTCWT can provide useful characterization of image structure, multiresolution, sparse representation and a high degree of shift-invariance in its magnitude, which make DTCWT superior to DWT in preserving details. DTCWT can also give phase information which plays an important role in medicine image. In addition, there is exactly an interval delay for sampling between the two trees, which makes the sample extracted by b-tree is exactly the sample value that a-tree discards, so the number of decomposition directions can be increased while ensuring translation shift-invariance.

2.2.2 NSCT

Since the two-dimensional wavelet transform can only capture the information in the directions of 0° , 45° and 90° , and can not represent the curve optimally, it can only approximately represent it by points. Do et al. proposed contourlet transformation which can fit a smooth curve with coefficients less than wavelets. However, this transform still uses downsampling and has shift-variant. NSCT is a redundant transform which is shift invariant in nature and provides rich directional information. This directionality is useful in the reconstruction of images. NSCT also gives a number of subbands on decomposition; therefore, it provides flexibility in image fusion. Two shift invariant filter banks, namely, nonsubsamped pyramid (NSP) and nonsubsamped directional filter banks (NSDFB) have been used in the construction of NSCT. Multiscale property has

been provided by NSP whereas directional information is obtained by NSDFB. The NSCT filter bank structure is shown in Fig. 3. The input image is first decomposed by NSDFB to obtain different frequency sub-bands, and then decomposed by NSDFB to obtain different direction sub-bands. After K levels of decomposition, $K + 1$ sub-band images with the same size as the source image are generated, namely, a low-frequency image and K high-frequency images are generated. In this way, the problem of shift-variant is well solved, but the algorithm has considerable computation complexity.

2.2.3 NSST

The NSST is an extension of the NSCT in multidimensional and multidirectional case that combines the multi-scale and direction analysis, separately. Firstly, the NSLP is used to decompose an image into low and high-frequency components, and then direction filtering is employed to get the different subbands and different direction shear let coefficients. The direction filtering is achieved using the shear matrix, which provides various directions. Once different frequency sub-bands are generated by decomposing, shearing filters is used for direction decomposition instead of the contourlet filter. The advantage is that there is no limit to the number of directions and size of the support base, but the computational efficiency is very high.

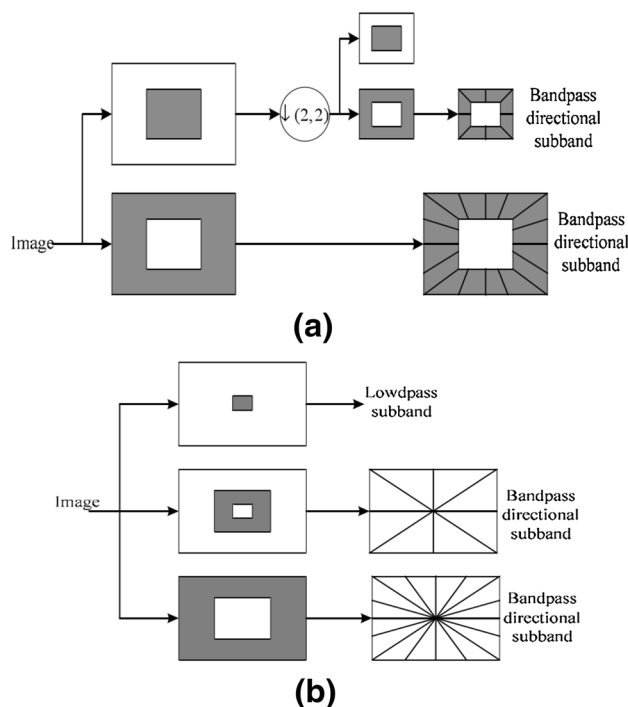


Fig. 3 Multi-scale image transformation. a Decoruposition fraruework of CT. b Decoruposition fraruework of NSCT

3 Our proposed fusion algorithm

3.1 General idea

In order to analyze our proposed fusion algorithm based on improved Deep Convolutional Neural Network (DSNN), the general idea is shown in Fig. 4. Now let’s briefly introduce each module.

- (1) The back propagation algorithm is adopted as basic training unit.
- (2) Multiple well-trained basic units are stacked to form DSCNN, and the parameters of the whole network are adjusted finely by means of end-to-end.
- (3) Image A and B that are ready for fusion, are respectively decomposed into their own high-frequency and low-frequency images through the same DSCNN.
- (4) The high-frequency and low-frequency images of A and B are fused by corresponding fusion rules to obtain the fused high-frequency and low-frequency images.
- (5) The results in (4) are put back into DSCNN to reconstruct and get the final fusion image.

3.2 Construction and training for improved CNN

3.2.1 The structure of our CNN

Our improved CNN is stacked by many basic units, where the basic unit is composed of high frequency and low frequency subnets. The high frequency and low frequency subnets are respectively composed of three convolution layers, wherein the first layer limits the input information; the second Layer is to combine the information; the third layer is to merge this information into an image of high-frequency and low-frequency. The structure and the construction method is as follows:

- (1) Feature mapping of H1 layer is obtained as follows:

$$I_{H1}^i = f(\omega_i \otimes x + \theta_i) \tag{1}$$

where \otimes represents the convolution operation, f is the activation function, ω_i is the i th convolution kernel, $i = \{1, 2, \dots, n_1\}$, n_1 is the number of feature mappings of H1 layer, and θ_i indicates the bias. Similarly, the feature mapping I_{L1} of L1 layer, can be obtained.

- (2) The feature mapping of H2 layer is obtained as follows:

$$I_{H2}^j = f\left(\sum_i^{n_1} \omega \otimes I_{H1}^i + \theta_j\right) \tag{2}$$

where I_{H2}^j is the j ($j = \{1, 2, \dots, n_2\}$)th result for I_{H2} . n_2 is the number of feature maps in H_2 layer, θ_j represents the j th bias. Similarly, the feature mapping I_{L2} of layer $L2$ can be obtained.

- (3) I_{H2} is used to replace I_{H1} in Eq. (2), the high-frequency image I_{H3} can be obtained. Similarly, the low-frequency image I_{L3} can be obtained.
- (4) The high frequency image I_{H3} and the low frequency image I_{L3} are convoluted to obtain the reconstructed image y .

3.2.2 The training process of our CNN

DSCNN training includes the basic unit training and the stacked network training, whose detail is shown as follows:

3.2.2.1 Basic unit training The convolution kernel from layer X to layer H1 is initialized into a Gaussian Laplace filter; and the convolution kernel from layer X to layer L1 is initialized to a Gaussian filter. The rest of the convolution kernels are initialized by The K method proposed in [18]. All the network biases are initialized to 0.

The unsupervised training is used to train the basic unit. Enter the training data $\{x_s, y_s\}_{s=1}^N$, where $y_s = x_s$. Suppose the output of the network is z_s , the loss calculation result

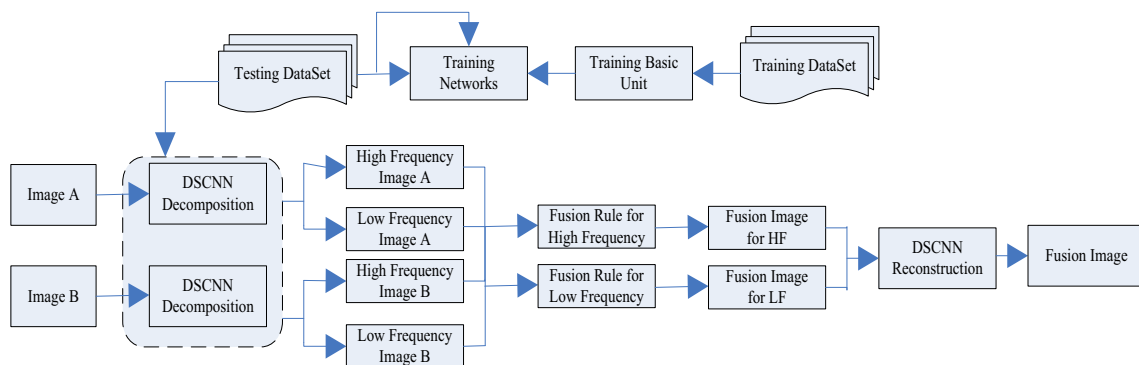


Fig. 4 General idea and diagram

will be too large when the common square error is used as the objective function, which will affect the stability of the network. Therefore, we define the objective function as follows [19]:

$$\Theta(W, \theta) = \frac{1}{Nmn} \sum_{s=1}^N \sum_{u=1}^m \sum_{v=1}^n (z_{iuv} - x_{iuv})^2 \tag{3}$$

where W is the convolution kernel, θ is the bias, N is the size of the training set, S is the current training sample, m and n are the size of a single image, z_{iuv} represents the value of a point of output results of the current sample. In this formula, the total error is averaged to each pixel, and the obtained value is small, which is convenient for training. Then the network is trained by back propagation algorithm.

3.2.2.2 Stacked network training Multiple trained basic units are connected by end-to-end way to form a stacked network. Then use the data set and objective function that are the same as the basic unit to adjust the whole network by means of End-to-End at the same time. Finally, get the stacked convolution neural network.

After the above training, our proposed model can decompose the image into high frequency and low frequency images, and realize the reconstruction within the range of error expected. It can show the process of our model when the number of stacks is 3, where the actual size of convolution kernel is 5×5 . For the convenience of observation, 4 convolution kernels in the first layers of the high-frequency subnets and low-frequency subnets are magnified by 10 times, and the visualization result is shown in experimental section. The first image shows the convolution kernel initial state (4 convolution kernels are the same), and the remaining four images show the convolution kernel state after training [20, 21].

3.3 Fusion image with proposed model

The network structure of the proposed CNN-based fused image is shown as follows:

1. Source images A and B are input into the well-trained CNN model, and get high frequency images A_{H3} and B_{H3} and low frequency images A_{L3} and B_{L3} .
2. The high-frequency information mainly corresponds to the gray abrupt change region such as the edge contour. In order to select brighter pixels to form the edge, most researches adopt the high-frequency fusion rule that the selected corresponding coefficient should be larger. Human eyes are sensitive to local information, so using local parameters to determine the fusion rules is more in line with human visual characteristics. The common indicators to measure the abundance of

local information include local variance, local entropy, local roughness, etc. The larger the value is, the more abundant the local information is. In addition, the local variance is the simplest and the most commonly used, and the larger the value is, the more abundant the information of the target edge and the details of the region. Therefore, the fusion rule is to choose the largest local variance [22]:

$$F_{H3}(x, y) = \begin{cases} A_{H3}(x, y), & \sigma_{A_{H3}}(x, y) > \sigma_{B_{H3}}(x, y) \\ B_{H3}(x, y), & \text{else} \end{cases} \tag{4}$$

where $\sigma(x, y)$ represents the local variance of the point (x, y) .

3. The common low-frequency fusion rule is based on the weighted average of the gray value or local parameters, or to select the greater value, etc. Considering that the local energy can reflect saliency of the target locality, the larger the value is, the more prominent the target is; and the local matching degree can be used to measure the similarity of the locality of the images to be fused. Therefore, combining these two parameters can be adaptive to determine the weight coefficient. The specific method is as follows:

- (1) The local energy of low-frequency image $A_{L3}(x, y)$ is defined as:

$$E^{A_{L3}}(x, y) = \sum_{q=-Q}^Q \sum_{p=-P}^P [A_{L3}(x + p, y + q)]^2 \tag{5}$$

where P and Q represent the size of the local window of the control point (x, y) , $A_{L3}(x + p, y + q)$ indicates the value A_{L3} of the point $(x + p, y + q)$. Similarly, the local energy of B_{L3} can be obtained as $E^{B_{L3}}(x, y)$.

- (2) The matching degree of low-frequency images $A_{L3}(x, y)$ and $B_{L3}(x, y)$ with the corresponding areas:

$$MAB(x, y) = \frac{\sum_{q=-Q}^Q \sum_{p=-P}^P A_{L3}(x + p, y + q) B_{L3}(x + p, y + q)}{E^{A_{L3}}(x, y) + E^{B_{L3}}(x, y)} \tag{6}$$

The matching degree reflects the correlation between A_{L3} and B_{L3} , if $A_{L3} = B_{L3}$ in the corresponding area, then $MAB(x, y) = 1$, the information matching degree in this area is the highest.

- (3) Low frequency images fusion based on saliency metrics and matching degree [23, 24]. Suppose α to be the threshold of matching degree, $MAB(x, y) < \alpha$ indicating that there is significant difference between the corresponding regions, a consolidation rule to select the larger

value should be adopted, as shown in Eq. (7). Otherwise, if $MAB(x, y) \leq a$, indicating a little difference between the two images, the adaptive weighted averaging method is adopted [25]. The method is shown in Eq. (8).

$$F_{L3}(x, y) = \begin{cases} A_{L3}(x, y), & E^{A_{L3}}(x, y) > E^{B_{L3}}(x, y) \\ B_{L3}(x, y), & \text{else} \end{cases} \tag{7}$$

$$F_{L3}(x, y) = \begin{cases} \omega_L A_{L3}(x, y) + \omega_S B_{L3}(x, y), & E^{A_{L3}}(x, y) > E^{B_{L3}}(x, y) \\ \omega_S A_{L3}(x, y) + \omega_L B_{L3}(x, y), & \text{else} \end{cases} \tag{8}$$

where ω_L and ω_S represent the larger weight and the smaller weight, respectively, and

$$\omega_L = \frac{1}{2} + \frac{1}{2} \left(\frac{1 - MAB(x, y)}{1 - a} \right) \tag{9}$$

$$\omega_S = 1 - \omega_L \tag{10}$$

- (4) F_{H3} and F_{L3} are input into the network, get the fusion result F.

4 Experimental results and analysis

4.1 Parameters setup and selection

The simulations are all implemented by MATLAB (R2010a) on a personal computer with 2-GB memory, 2.94-GHz Intel Core i5-7500 processor, GPU is NVIDIA Tesla V10. Training environment is Lasagne0.2; Python version is 2.8. All experimental data is from Public Hospital, which is real and reliable, and is approved by the medical experts.

There are two main factors that affect the effect of image fusion: learning ability and the fusion rule, where learning ability is mainly affected by the dataset used in training, the initialization of convolution kernel, the number of convolution kernels of each layer in each basic unit (n_1, n_2), the learning rate, the number of stacked basic units, and the activation function f . The parameters of the fusion rule include the window size when the largest high-frequency variance is selected, the matching threshold a , and the window size for calculating the local energy.

4.1.1 Impact and selection of data sets

The fusing image objects include multi-band images, multi-modal images, etc. In order to facilitate the training, all images are cut and scaled and normalized. In addition, extract another 200 images for validation set, and the

proportion of each image type is the same as that of the training set.

4.1.2 Influence and selection of convolution kernel initialization

Applying He K method to initialize the convolution kernel can improve the stability and generalization ability of the network. However, the network initialized only by this method cannot classify and decompose the image. Therefore, in order to ensure that the network can obtain the high-frequency information and the low-frequency information respectively, Gaussian Laplacian filter and Gaussian filter are used to initialize the first-layer filters for the high-frequency subnet and the low-frequency subnet. Figure 8 shows the different effect of two initializations on network training.

It can be seen from Fig. 5 that HeK algorithm is used for initialization both in the proposed method and all layers. No matter whether the network is stacked or not, the errors of these two are decreased steadily. Although the final error of the proposed method is slightly larger than He K initialization method, the proposed method can decompose the image information, which is more conducive to the subsequent fusion process.

4.1.3 The influence and selection of the number of basic unit convolution kernels

Generally, the larger the number of convolution kernels is, the stronger the adaptability of the network is, and the easier for the image feature to be extracted. However, the downsampling layer is removed in our CNN model. The size of each convolution kernel is the same as output. An excessive number of convolution kernels results in increased memory space and computational complexity, therefore, on the premise to ensure the accuracy of feature

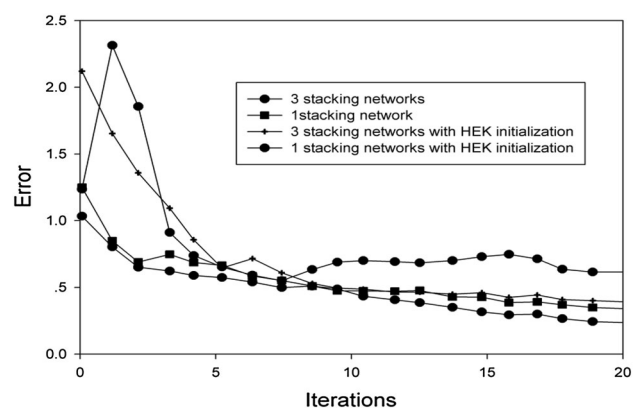


Fig. 5 Influence for different initialization networks

extraction, the number of convolution kernels should be as little as possible.

Figure 6 shows the training results for basic unit with different numbers of convolution kernels. The two Arabic numerals in the figure respectively denote the values of n_1 and n_2 . It can be seen from the figure that when $n_1 = n_2$, the overall network error decreases more steadily and the final error is the smallest when $n_1 = n_2 = 4$. Therefore, the scale of the network is selected as: $n_1 = n_2 = 4$.

4.1.4 Impact and selection of learning factor

The learning factor also has a great influence on the speed and stability of network training. Excessive learning rate will make the network difficult to converge; if the learning rate is too low, the speed of training will be declined. While training the basic unit, the network error is large, and the layers are shallow. Taking a larger learning rate helps to improve network training speed, the learning rate is set to 0.01; when the error is stable, well-trained basic units are stacked up to train the entire stacking network. With the increase of the number of stacks, the difficulty of training gradually increases. Therefore, the learning rate needs to be reduced so as to obtain an ideal training result. Figure 7 shows the influence of different learning rates on training when the number of stacks is 3. Apparently, the greater and more unstable the learning rate is 0.01, the larger is the final error, therefore, in this article, the learning rate is selected to 0.001 (Fig. 8).

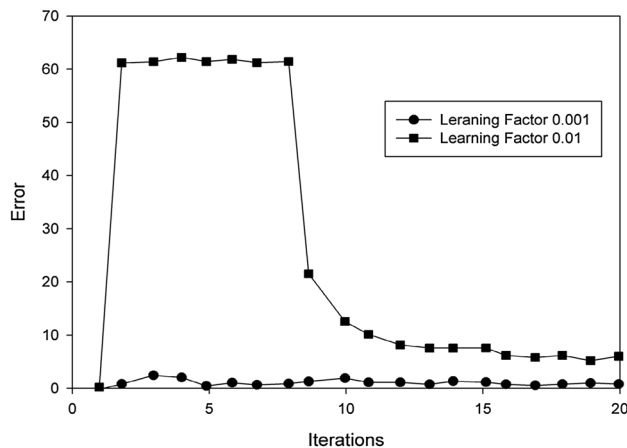


Fig. 7 Influence of our modal form different learning factor

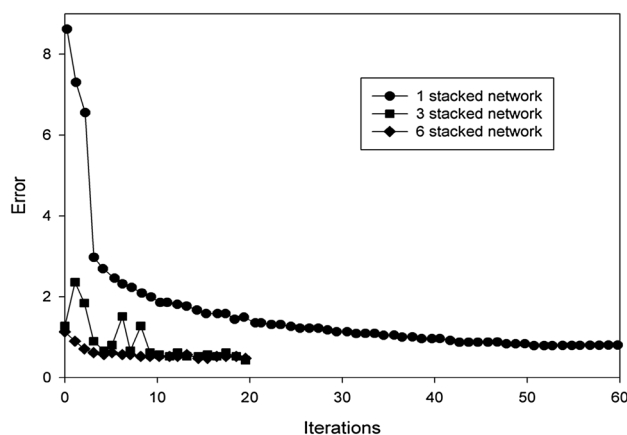


Fig. 8 Influence of our network form the number of basic units

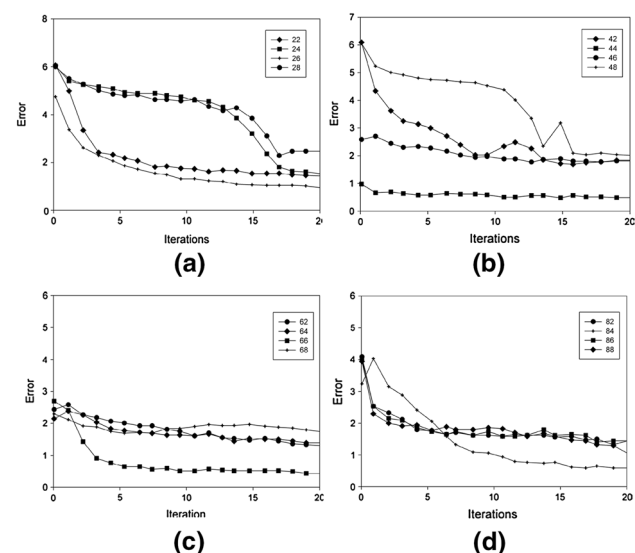


Fig. 6 Influence from the number of convolution kernels. a $n_1 = 2$, b $n_1 = 4$, c $n_1 = 6$, d $n_1 = 8$

4.1.5 Impact and selection of the number of stacks

Stacked network is to improve the accuracy of image decomposition. It is generally considered that, with the same error, the larger the number of stacks, the more detailed the network image decomposition is simulation testing shows the network error when the network has different numbers of stacks. Obviously, when the number of stacks increases to 6, the network error decreases more steadily. However, experiments also show that with the increase of the number of stacks, the reconstructing ability of our model to the detailed information like the edge information worsens, and false edges appear in some details. The experiments show the network with 1 stack, 3 stacks and 6 stacks, respectively, the high-frequency images, low-frequency images and reconstructed images when our proposed improved CNN are trained to be with a stable error. It can be seen that in the high frequency image, the corresponding area to the upper right corner rectangle is the clearest with 6 stacks, but is the most obscure with 1 stack. However, as for the corresponding

position to the dashed line of the reconstructed image, when the number of stacks is 6, false edges appear in the areas like the vessel area. Overall, when the number of stacks is 3, both the reconstruction effect and the training time of the images are ideal. Therefore, the number of stacks is set to 3.

4.1.6 The influence and selection of activation function

Activation function is another important factor affecting network learning ability and the training speed. The commonly used activation functions include Sigmoid, Tanh-ReLU, etc. Although ReLU function has proved to be more conducive to network training, but its output is unlimited. When the output is converted to images, there will be error. The outputs of Sigmoid and Tanh are closed, however, compared to Sigmoid, Tanh function is difficult to saturate, and more conducive to training. Therefore, in this network, Tanh is selected as the activation function.

4.1.7 Influence and selection of other parameters

Other parameters include the local variance window size in the high frequency fusion, the local energy window size and the matching degree of the low frequency fusion. The size of local window is usually selected to 3×3 , 5×5 , etc. The larger the window is, the more difficult for the local information to be described. Too small window is difficult to reflect the local features. Therefore, in order to compromise it, the size of the local window is selected to be 5×5 . The larger the matching threshold, the lower the probability that the information is classified as a match, the more results that will be determined by the degree of salience, according to the test, the matching degree is selected as $a = 0.85$.

4.2 Experimental results and analysis

A number of illustrative examples are presented to assess the effectiveness of the proposed image fusion method. The fusion results obtained by the proposed method on the different datasets of the CT and MR images are presented. Moreover, the results produced by the proposed method are also compared with the several existing fusion methods in the qualitative and quantitative manner. The essential requirement of this algorithm is that the input images should be preregistered. All images used for testing were of size 256×256 and downloaded from the Harvard University site (<http://www.med.harvard.edu/AANLIB/home.html>). Any fusion algorithm in addition to subjective analysis should validate the system using some objective metrics.

The proposed fusion method is compared with six representative multi-focus image fusion methods, which are the NSST [5], NSCT [4] and DTCWT [3]. Due to fusion rules have obvious effects on the fusion results, literature [6] compared the fusion results of the NSCT and DTCWT under the condition of different decomposition levels and different fusion rules: big absolute for high-frequency, average for low-frequency. Therefore, we give the results for the fusion rules in this paper, denoted as ACNN. At the same time, in order to compare the influence of the convolution kernels in different initialization, the fusion rule is set as the big local variance for high-frequency and the regional matching degree-merge for low-frequency, so our proposed method and deep stack network adopt He K initialization method to get fusion result, denoted as HCNN.

The decomposition layers of NSST and NSCT are set up as 4, whose decomposition direction are $\{6,10,10,18\}, \{4,8,8,16\}$, respectively. The filters are ‘maxflat’ and ‘pyrexc’ respectively; The DTCWT decomposition layer is also 4, and the decomposition filter in the first layer selects ‘5–3’ and the filter for remaining layers are selected ‘q-6’. Each group images are followed by two source images, NSST-maxflat (NSSTM) fusion result, NSST-pyrexc (NSSTP) fusion result, NSCT-maxflat (NSCTm) fusion results, NSCT-pyrexc fusion (NSCTP)} DTCWT fusion results, ACNN fusion result, HCNN fusion results, the fusion result.

In general, the six methods preferably colligate the difference information of source image. Compared with the multi-scale transform method, our proposed algorithm has clearer result, such as the vessel region in brain image within bigger contrast and clearer edge. In order to make it clear and clean, it is easier to identify the target. In Fig. 9, there is a halo on the edges, such as NSCT, NSST and DTCWT. The boundary of the fusion result is clearer. In Fig. 10, There is no artifact in our result, and the texture is sharper. ACNN is superior to the multi-scale method because it uses more filters and after extensive sample learning make the filter can seem to be adaptive. While compared with the fusion results of HCNN and ACNN, our fusion results have higher contrast and clarity.

Because difference of subjective observation exists among different people, the following objective indicators that are commonly used in the literature are selected to measure the fusion method. Considering the purpose of image fusion is mainly in three aspects of integrating different information of source image, improving the sharpness and visual perception quality, and the fact that there is no uniform standard to measure the image quality, the following three types of indicators are selected to evaluate the fusion results: The first category which describes the abundance degree of the image includes Standard

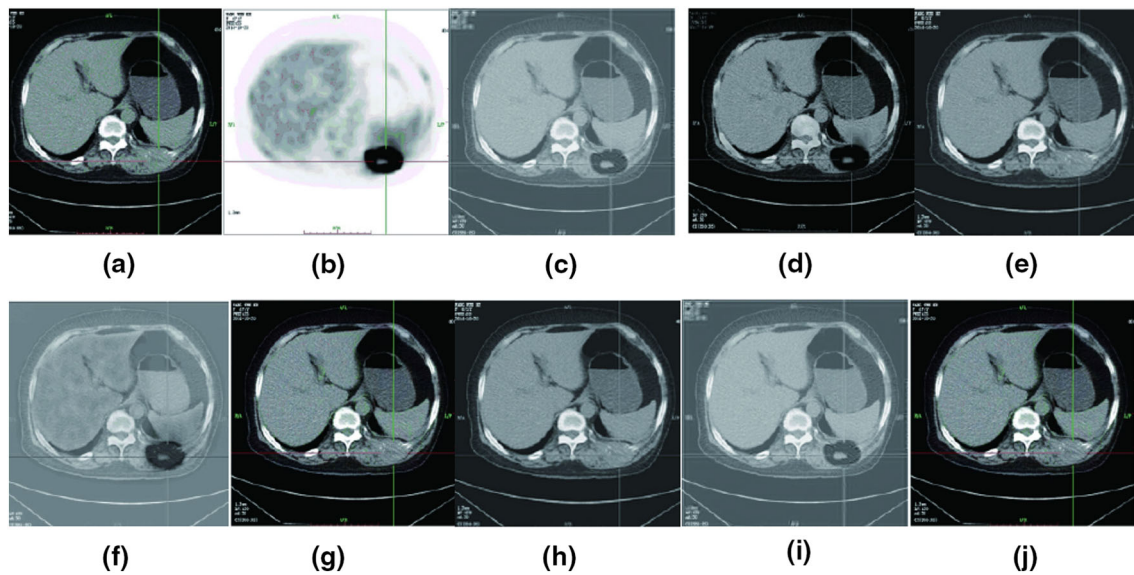


Fig. 9 Fusion results for different algorithms. **a** CT, **b** PET, **c** NSSTM, **d** NSSTP, **e** NSCTM, **f** NSCTP, **g** DTCWT, **h** ACNN, **i** HCNN, **j** proposed

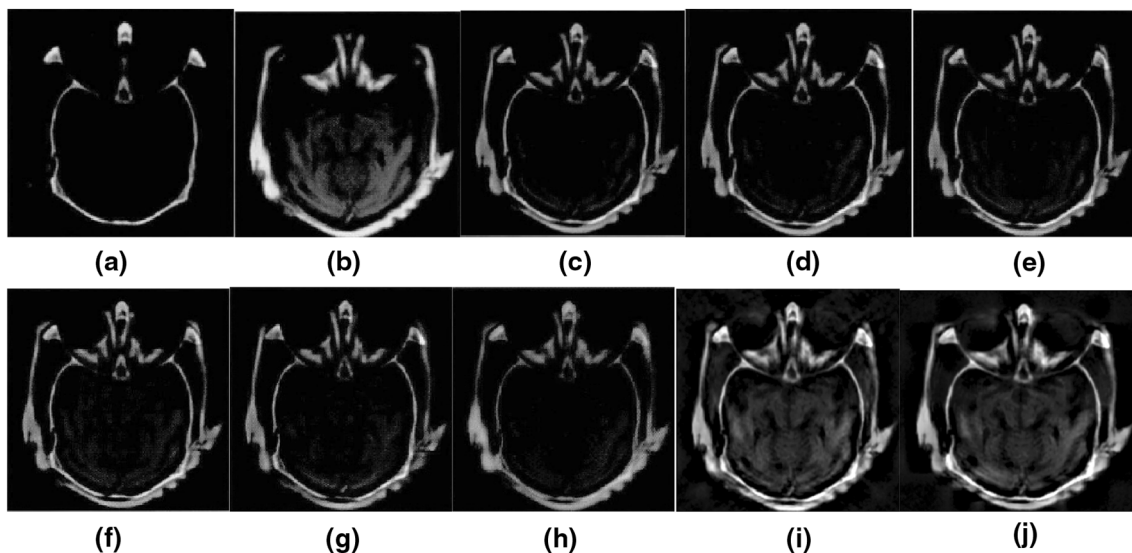


Fig. 10 Fusion results for different algorithms. **a** CT, **b** MRI, **c** NSSTM, **d** NSSTP, **e** NSCTM, **f** NSCTP, **g** DTCWT, **h** ACNN, **i** HCNN, **j** proposed

Deviation (SD), Entropy (E) and Mutual Information (MI); the second category which is to measure the sharpness of the image include Contrast (C), Average Gradient (AU) and Spatial Frequency (SF); the third category which reflects the visual quality includes Uniform Image Quality Indicators (UIQI). As for all these indexes, the larger the value is, the better the fusion effect will be. It should be noted that the universal image quality indicators need to be compared with the ideal reference image, but the ideal reference image is difficult to acquire, therefore, in the proposed method in this article, two source images are compared first and then obtain the average. In addition, the

time to fuse is given in the algorithm, the specific results shown in Table 1. It can be seen from Table 1 that, among 8 groups of indicators, 8 optimal items are the fusion results of NSCT-P; 3 optimal items are the fusion results of DTCWT; but the number of optimal items of our fusion results is up to 27. In the automatic target recognition, the fusion speed is another important indicator, DTCWT has the shortest running time, followed by the proposed method, and the slowest is NSCT.

Our proposed fusion method obviously outperforms the other fusion methods. The proposed fusion scheme achieves 62.76%, 28.56% higher entropy than

Table 1 Quantitative analysis for different algorithms

No	Modal	Size	Methods	SD	E	MI	C	AG	SF	UIQI	Time (s)
1	CT and MRI (abdomen)	256 × 256	NSSTM	46.565	7.165	6.148	33.676	5.354	12.202	0.609	9.773
			NSSTP	46.670	7.165	6.116	33.982	5.335	12.465	0.608	9.893
			NSSCM	46.568	7.199	6.052	33.767	5.361	12.207	0.609	26.692
			NSSCP	46.674	7.167	6.119	33.763	5.341	12.172	0.611	27.835
			DTCWT	46.323	7.170	6.062	33.564	5.192	11.955	0.610	0.076
			ACNN	45.806	7.168	5.183	33.856	7.084	13.864	0.624	0.366
			HCNN	45.818	7.214	6.076	33.779	5.193	11.321	0.624	3.101
			Proposed	45.907	7.176	5.147	33.973	7.112	13.640	0.611	3.115
2	CT and PET	464 × 464	NSSTM	69.889	7.612	7.261	61.221	4.697	8.469	0.644	25.045
			NSSTP	69.904	7.645	7.284	61.395	4.635	8.358	0.651	24.163
			NSSCM	69.903	7.560	7.230	60.699	4.680	8.433	0.611	25.116
			NSSCP	69.947	7.599	7.299	60.682	4.627	8.340	0.654	71.369
			DTCWT	71.499	7.589	7.223	60.694	4.220	7.918	0.694	0.6982
			ACNN	69.823	7.587	6.335	62.126	6.764	11.268	0.643	1.222
			HCNN	69.824	7.538	7.331	61.367	5.699	6.954	0.647	9.009
			Proposed	75.422	7.622	6.344	64.426	6.545	11.354	0.632	8.763
3	CT and MRI (Brain)	512 × 512	NSSTM	21.963	6.271	3.545	12.554	2.961	8.916	0.846	32.264
			NSSTP	22.091	6.274	3.648	12.443	2.959	8.932	0.832	34.025
			NSSCM	22.034	6.298	3.557	12.468	2.991	8.906	0.846	91.365
			NSSCP	22.201	6.265	3.469	12.484	2.968	8.658	0.846	84.261
			DTCWT	21.658	6.187	3.477	12.567	2.854	8.958	8.841	1.269
			ACNN	20.374	6.298	3.532	11.586	3.156	8.813	0.850	1.438
			HCNN	20.943	6.287	3.721	12.819	2.278	6.930	0.951	11.125
			Proposed	21.386	6.188	3.464	11.503	3.395	8.031	0.832	11.046

source CT and MR images, respectively. Moreover, the proposed method gains approximately 6.62, 6.07, 5.1, 4.56, 3.74, 2.08, 1.16 and 3.14% higher entropy than others methods, respectively. These results ensure that the more information lies in the fused images obtained by the produced method than others. The larger value of the MI and SF metric of the fused images produced by the proposed method assures the more information preservation and more activity and clarity level in the fused images. It gains approx 11.73, 7–11 and 4–5% higher SF values than the DTCWT, NSCT and NSST based fusion method, respectively. Moreover, it has 26.9% and 1.44–26.21% larger MI values than the HCNN and ACNN based fusion method, respectively. Hence, on the basis of quantitative and visual analysis of results, it is observed that the proposed fusion algorithm outperforms the others by producing good quality of fused images with more details and edge information present in the source image.

5 Conclusion

The deep stacked neural network fusion method proposed in this paper consists of three parts: remove the down-sampled layer of the traditional convolutional neural network, initialize the first layer network convolution kernel with Gauss-Laplace filter and Gaussian filter, then use HeK-based method to initialize the convolution kernel of the rest layers, construct the basic unit, and use the back propagation algorithm to train the basic unit; (2) Train multiple basic units that are sacked with the thought of SAE to get the deep stacking neural network; (3) Use this stacking network to decompose the input images to obtain their own high frequency and low frequency images, and combine the rule of selecting the largest local variance and the rule of regional matching to fuse the two high frequency and low frequency images, and put the fused high frequency and low frequency images back to the last layer of the network to get the final fusion images. The proposed method can adaptively decompose and reconstruct the image in the fusion, only one high frequency and one low frequency image are needed, there is no need to define the

number of filters and the filter type manually, and there is no need to select the number of decomposition layers and the number of filtering directions, which can greatly solve the problem that the fusion algorithm depends greatly on the prior knowledge. The results show, overall, the proposed method has better effect than DTCWT, NSCT and NSST that have excellent fusion performance; although the fusion speed cannot surpass DTCWT, the speed of the proposed method is much faster than that of NSCT and NSST which have good fusion quality. It should be noted that although this method can decompose and reconstruct the image adaptively, the fusion rule still need to be defined manually. Therefore, the next step we will focus on improving the fusion rule so as to realize the adaptive multi-modal image fusion.

References

- Shu-Tao, Li, Xu-Dong, Kang, Leyuan, Fang, et al.: Pixel-level image Fusion; a survey Fusion of the state of the art. *Inf. Fusion* **33**, 100–112 (2017)
- Li, H., Manjunath, B., Multisensor, S.M.: Image Fusion using for the wavelet transform. *Graph Models Image Process* **57**(3), 235–245 (1995)
- Lewis, J.J., Callaghan, R.J., Nikolov, S.G., et al.: Pixel-and region-based image fusion with complex wavelets. *Inf. Fusion* **8**(2), 119–130 (2007)
- Liu, Y., Liu, S., Wang, Z.: A general framework for image fusion based on multiresolution transform and sparse representation. *Inf. Fusion* **24**, 147–164 (2015)
- Easley, U., Labate, D., Lim, W.Q.: Sparse directional image representations using the discrete shearlet transform. *Appl. Comput. Harmon. Anal.* **25**(1), 25–46 (2008)
- Li, S., Yang, B., Hu, J.: Performance comparison of different multi-resolution transforms for image fusion. *Inf. Fusion* **12**(2), 74–84 (2011)
- Yan-Ming, Guo, Liu, Yu., Oerlemans, A., et al.: Deep learning for visual understanding; a review. *Neurocomputing* **187**, 27–48 (2016)
- Hong, L., Fang, L., Shu-Yuan, Y., et al.: Remote sensing image fusion based on deep support value learning networks. *Chin. J. Comput.* **39**(8), 1583–1596 (2016)
- Fan, L., Ze-Hua, C., Jing, C.: A new multi-Locus image fusion method based on deep neural network model. *J. Shandong Univ. (Eng. Sci.)* **46**(3), 7–13 (2016)
- Ng, W.W.Y., Zeng, U., Zhang, J., et al.: Dual autoencoders features for imbalance classification problem. *Pattern Recognit.* **60**, 875–889 (2016)
- Gehring, J., Miao, Y., Metze, F., et al.: Extracting deep bottleneck features using stacked auto-encoders. In: *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. Vancouver, Canada, pp. 3377–3381 (2013)
- Zhao, Z., Jiao, I., Zhao, J., et al.: Discriminant deep belief network for high-resolution SAR image classification. *Pattern Recognit.* **61**, 686–701 (2017)
- Krizhevsky, A., Sutskever, I., Hinton, U.E.: Imagenet classification with deep convolutional neural networks. In: *Proceedings of the Advances in Neural Information Processing Systems*. Nevada, USA, pp. 1097–1105 (2012)
- Zabalza, J., Ren, J., Zheng, J., et al.: Novel segmented stacked autoencoder for effective dimensionality reduction and feature extraction in hyperspectral imaging. *Neurocomputing* **185**, 1–10 (2016)
- Jiao, L.-C., Yang, S.-Y., Liu, F., et al.: Seven years beyond neural networks: retrospect and prospect. *Chin. J. Comput.* **39**(8), 1697–1716 (2016)
- Kingsbury, N.: A dual-tree complex wavelet transform with improved orthogonality and symmetry properties. In: *Proceedings of the International Conference on Image Processing*. Vancouver, Canada, vol. 12, no. 2, pp. 375–378 (2000)
- Do, M.N., Vetterli, M.: The contourlet transform: an efficient directional multiresolution image representation. *IEEE Trans. Image Process.* **14**(12), 2091–2106 (2005)
- He, K., Zhang, X., Ren, S., et al.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: *Proceedings of the IEEE International Conference on Computer Vision*. Santiago, Chile, pp. 1026–1034 (2015)
- Russakovsky, O., Deng, J., Su, H., et al.: Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**(3), 211–262 (2015)
- Bengio, Y.: Practical recommendations for gradient-based training of deep architectures. In: *Proceedings of the Neural Networks: Tricks of the Trade*. Berlin, Germany, pp. 437–478 (2012)
- Shi, J., Zhou, S., Liu, X., et al.: Stacked deep polynomial network based representation learning for tumor classification with small ultrasound image dataset. *Neurocomputing* **19**(4), 87–94 (2016)
- He, K.-M., Zhang, X.-Y., Ren, S., et al.: Deep residual learning for image recognition. [arXiv:1512.03385](https://arxiv.org/abs/1512.03385) (2015)
- Kong, W., Wang, B., Lei, Y.: Technique for infrared and visible image fusion based on non-subsampled shearlet transform and spiking cortical model. *Infrared Phys. Technol.* **71**, 87–98 (2015)
- Smith, E.P.U., Pham, I.T., Venzor, U.M., et al.: HgCdTe focal plane arrays for dual-color mid-and long-wavelength infrared detection. *J. Electron. Mater.* **33**(6), 509–516 (2004)
- Jagalingam, P., Hegde, A.V.: A review of quality metrics for fused image. *Aquat. Proc.* **4**, 133–142 (2015)



Kai-jian Xia was born in 1983, Jiangsu province, China. He got his master degree from Jiangnan University, and now, he's studying for a Ph.D. degree at the China University of Mining and Technology. His research direction is medical information and medical image proceeding and so on. Now he worked in Department of Computer, Changshu No. 1 People's Hospital.



Hong-sheng Yin born in March 1967, Ph.D., professor. It has long been engaged in the teaching and research work of mine monitoring and monitoring, information processing and computer network communication.



Jian-qiang Wang was born in Jiangsu province, China. He is vice director of Changshu No. 1 People's Hospital. His research direction is Hospital management and medical information.