Optik - International Journal for Light and Electron Optics 176 (2019) 567–578

Contents lists available at ScienceDirect

Optik

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

Original research article

Multi-focus image fusion using deep support value convolutional neural network

ChaoBen Du[a,*], SheSheng Gao[a], Ying Liu[b,c], BingBing Gao[a]

[a] School of Automation, Northwestern Polytechnical University, Xi'an, 710129, China
[b] Center for Image and Information Processing, Xi'an University of Posts and Telecommunications, Xi'an, 710121, China
[c] Key Laboratory of Electronic Information Application Technology for Scene Investigation, Ministry of Public Security, Xi'an, 710121, China

ARTICLE INFO

Keywords:
Multi-focus image
Convolutional neural network
Image fusion
Decision map

ABSTRACT

A novel multi-focus image fusion algorithm based on deep support value convolutional neural network (DSVCNN) is proposed for multi-focus image fusion. First, a deep support value training network is presented by replacing the empirical risk minimization-based loss function by a loss function based on structural risk minimization during the training of convolutional neural network (CNN). Then, to avoid the loss of information, max-pooling/subsampling of the feature mapping layer of a conventional convolutional neural network, which is employed in all conventional CNN frameworks to reduce the dimensionality of the feature map, is replaced by standard convolutional layers with a stride of two. The experimental results demonstrate that the suggested DSVCNN-based method is competitive with current state-of-the-art approaches and superior to those that use traditional CNN methods.

1. Introduction

In natural images, the imaging equipment usually captures a target image, which includes all the image objects that are effectively captured in focus. In general, by setting the focal length of the optical lens, only the objects in the depth-of-field (DOF) area are clearly visible in the picture, while others are blurred [1]. Consequently, due to the shortcomings of the depth-of-focus (DOF) of optical lenses in charge-coupled device devices, it is difficult to obtain an image where all the relevant objects are effectively captured in focus. To overcome this issue, an image fusion algorithm is introduced in this paper in which multiple source images of the same scene are combined to form a fused image where all the targets of interest are fully focused [2]. One of the basic requirements for image fusion is that all the details should be extracted from multiple source images and preserved in the final fusion image. To some extent, for multi-focus image fusion, only the focused regions in the multi-focus source images need to be preserved perfectly in the final fused image, while all the defocused regions should be completely removed [3].

A variety of multi-focus image fusion algorithms have been proposed over the last decade [2–8]. Overall, these methods can be classified into two categories: transform domain and spatial domain methods [2]. In the literature, multi-scale transform (MST) is one of the most popular transform domain methods [4]. Conventional MST image fusion methods include pyramid-based [6], wavelet transform-based [9], curvelet transform-based [10], shearlet transform-based [11] and non-subsampled contourlet transform-based (NSCT) [3,12] algorithms. Because representation approaches of these image are consistent with the human visual system (HVS); the transform domain methods are it generally considered highly effective in image fusion [13,14].

Spatial domain methods typically solve the fusion issue using pixel-wise gradient information [15–17] or image blocks [18–21];

* Corresponding author.
E-mail address: dcbxjdaxue@163.com (C. Du).

https://doi.org/10.1016/j.ijleo.2018.09.089
Received 24 February 2018; Received in revised form 3 June 2018; Accepted 17 September 2018

however, this approach often introduces many undesirable block artifacts [3]. In the last several years some block-based fusion methods have been published [20,22]. At present, the advanced pixel based fusion methods include guided filtering-based (GF) [23], image matting-based (IM) [15], dense scale invariant transform-based (DSIFT) [2], and homogeneity similarity-based (HS) [24] methods. These fusion algorithms perform well in extracting and preserving image detail.

In both transform domain and spatial domain image fusion algorithms, the decision map is a key factor in performing multi-focus image fusion. To improve the quality of the multi-focus image fusion result, the recently proposed image fusion methods have become increasingly more complicated. Over the last several years, both multi-focus image fusion methods and spatial domain-based methods have been widely introduced. The multi-focus image fusion methods, based on the simplest pixel, directly average the pixel values of all the input images. The direct averaging-based image fusion algorithms can rapidly and concisely obtain a fused image, which are their main advantage, but their disadvantage is that the fused image tends to produce a blurred effect, caused by the loss of some information from the source input images. To overcome these drawbacks, several promising pixel-based multi-focus image fusion methods have been proposed, including dense SIFT [2] and guided filtering [23]. The Dense SIFT and guided filtering-based methods produce a decision map by detecting focused pixels from each source input image and then extracting the clear area from every source input image based on an optimized decision map. The final fused image is produced by integrating the pixels in the clear areas from all the scenes. The decision map is used to identify the clear areas. The black areas of the decision map denote unfocused regions of a source image, while the white area of the decision map represents the clear region of a source image. The focused region is employed as the fusion decision map to guide the fusion process of the multi-focus image. This approach not only reduces the complexity of the procedure but also increases the reliability and robustness of the fusion image results. The multi-scale weighted gradient-based image fusion method presented in [25] reconstructs the fused image by making its gradient as close as possible to the magnitude of the merged gradient rather than employing a decision map. Although the new methods discussed above can obtain high-quality fused images, they can lose some of the source input image information as a result of inaccurate fusion decision maps.

Recently, a new spatial domain image fusion method was proposed, namely, the CNN-based image fusion method [26]. Although the CNN-based method has been widely applied in such fields as license plate recognition, face recognition, behavior recognition, image classification and speech recognition, it is seldom mentioned in the area of image fusion. Yu Liu first introduced CNN into multi-focus image fusion with satisfactory fusion results. The CNN-based algorithms can provide better performance than the traditional spatial transform-based algorithms. However, the CNN-based fusion algorithms in [26] have two drawbacks: In [26] the parameters of the filter in each layer are obtained through the minimization of empirical risk. However, using empirical risk minimization, it is difficult to ensure that the trained network will have good promotional performance. Max-pooling and subsampling reduces the resolution of the feature map, resulting in information loss. At the same time, there have also been some methods of combining support vector machines with CNN in recent years. However, these methods are mainly used in pattern recognition [27] and classification [28], and have not been applied to multi-focus image fusion. The method that appears at present just simply combines SVM and CNN, and does not have a network that makes both together.

In this article, a novel multi-focus DSVCNN-based image fusion method is presented to overcome the deficiencies of the CNN-based image fusion methods. We demonstrate that DSVCNN can successfully overcome the two problems described above. We demonstrate that the decision map produced by the DSVCNN is reliable and that it can obtain high quality image fusion results. The experimental results show that the proposed multi-focus image fusion method achieves state-of-the-art fusion performance in terms of both qualitative and quantitative evaluations. The contributions of this article are as follows:

First, the DSVCNN model is not dependent on the empirical risk in the learning process. It can adaptively learn the optimal support value filter at all levels of decomposition.

Second, the support value filter can find the essential characteristics of the image and effectively extract details from all levels of the image.

Third, to avoid information loss, the max-pooling and subsampling of the feature map layer of conventional CNN, which is employed in all conventional CNN frameworks for dimensionality reduction, is replaced by standard convolutional layers with a stride of two.

The remainder of this article is arranged as follows. The basic theoretical underpinnings of deep support value learning networks are introduced in Section 2. Sections 3 and 4 describe the proposed CNN method and the improved CNN method, respectively. The implementation of the proposed CNN method for multi-focus image fusion is presented in Section 5. A detailed discussion and conclusions based on the experiments are respectively presented in Sections 6 and Section 7.

2. Deep support value learning net works

All the models in the deep neural network include the CNNs; the filters in the layers are mostly obtained by minimizing the empirical risk. However, it is difficult to ensure that a trained network has a good generalization performance. Similar to the support vector machine (SVM), the basic network unit we used is shown in Fig. 1

Let $\mathbf{x} \in R^d (y \in R)$, where R^d denotes input space, y is the supervisor's response or output [29], d is the dimension. $\mathbf{x}$ undergoes the convolution operation C, and then, through the hidden layer of neurons, transfers the function mapping, that is, $\phi(\mathbf{x})$: $R^d \rightarrow R^q$ (where q represents the dimension of the feature space). Output occurs via the linear layer and then goes through the weight W and bias b. During the training of the basic unit, the training data $\{x_i, y_i\}_{i=1}^N$ is input to the basic unit; then, the estimated function of the output is as follows.

$$f(C, W, x_i) = \mathbf{W}^T \phi(\mathbf{x}) + b. \tag{1}$$

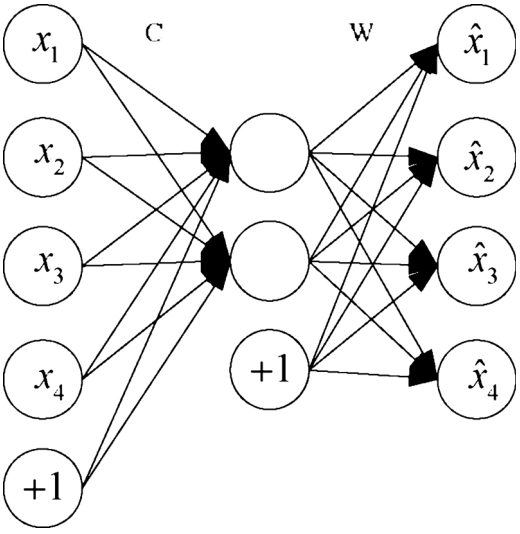


Fig. 1. Deep support value network learning basic unit.

The objective function can be defined as

$$R(C, W) = \sum_{i=1}^{N} L[y_i, f(C, W, x_i)] + \gamma \frac{||W||^2}{2} \tag{2}$$

where $L[y_i, f(C, W, x_i)] = [y_i - \mathbf{W}^T\phi(\mathbf{x}) - b]^2$.

Based on the idea of the deep neural learning network, we assume that the expected output is equal to the input. Similar to the idea of solving the weights in SVM, the estimation function of the basic network can be given as follows.

$$f(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i K(C\mathbf{x}, C\mathbf{x}_i) + b, \tag{3}$$

where α_i is the support value of the support vector $\mathbf{x}_i$, and $K(C\mathbf{x}, C\mathbf{x}_i) = \phi(C\mathbf{x})^T\phi(C\mathbf{x}_i)$, $i = 1, \cdots, N$ is a kernel function. The matrix form of the estimated function is

$$\begin{bmatrix} 0 & \mathbf{1}^T \\ \mathbf{1} & \Omega \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{Y} \end{bmatrix}, \tag{4}$$

where $\Omega_{i,j} = K_{i,j} + I_{i,j}/\gamma$, $K_{i,j} = K(C\mathbf{x}_i, C\mathbf{x}_j)$, $\mathbf{Y} = [y_i, \ldots, y_N]^T$, $\boldsymbol{\alpha} = [\alpha_i, \ldots, \alpha_N]^T$, and $\mathbf{1} = [1, \ldots, 1]^T$. Using the iterative algorithm to optimize the network parameters, the update process is as follows:

First, for a fixed C, the explicit solution of (4) is

$$\begin{cases} \alpha = \frac{\mathbf{1}^T\boldsymbol{\Omega}^{-1}\mathbf{Y}}{\mathbf{1}^T\boldsymbol{\Omega}^{-1}\mathbf{1}} \\ b = \boldsymbol{\Omega}^{-1}(\mathbf{Y} - b\mathbf{1}) \end{cases}. \tag{5}$$

In the next step, we assume that $A = \boldsymbol{\Omega}^{-1}$ and $B = \frac{\mathbf{1}^T\boldsymbol{\Omega}^{-1}}{\mathbf{1}^T\boldsymbol{\Omega}^{-1}\mathbf{1}}$. Then, (5) can be expressed as follows:

$$\begin{cases} \alpha = A(I - \mathbf{1}\mathbf{B}^T)\mathbf{Y} \\ b = \mathbf{B}^T\mathbf{Y} \end{cases}, \tag{6}$$

where $\mathbf{Q} = A(I - \mathbf{1}\mathbf{B}^T)$ is a matrix of $N \times N$. If the support value of the pixel (x, y) is approximated by the corresponding support value of the input vector in the mapped neighborhood center, we can obtain the support values C of the entire image by convolving the image with the support value filter derived from the central row vector of matrix Q [30].

3. CNN

CNN is a typical deep learning model that learns a hierarchical representation of an image at different abstraction levels [25]. From Fig. 2, we can see that an emblematic CNN model contains an input layer, convolutional layers, subsampling/max-pooling layers, a fully connected layer, and an output layer.

The input of the CNN is the original image X in most cases. In this article, the notation H_i represents the feature map of the $i-th$ layer of the CNN (where $H_0 = X$). We assume that H_i is a convolutional layer in the convolutional neural network; the generation of H_i is rewritten as follows:

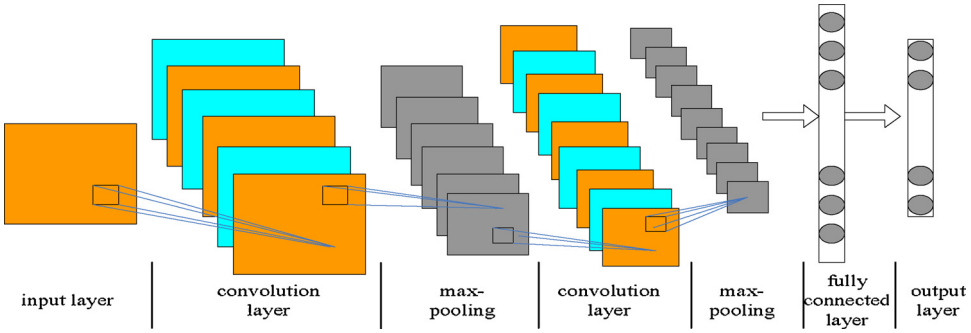


Fig. 2. Typical Structure of a CNN.

$$H_i = f(H_{i-1} \otimes W_i + b_i) \tag{7}$$

where W_i is the convolutional kernel, b_i is the bias, and $\otimes$ represents the convolutional operation. Here, $f(\bullet)$ is the non-linear ReLU activation function.

The max-pooling layer of the typical CNN closely follows the convolutional layer in most implementations; then, the feature map is obtained by the max-pooling layer according to a specific max-pooling rule. By alternating the multiple convolutional and max-pooling layers, the CNN depends on a fully connected network to classify the extracted features using the CNN framework to obtain the probability distribution based on the input. The residuals of the conventional CNN are propagated backward through the gradient descent method [31].

4. Improved CNN model

The proposed CNN model used in our experiment is different from other standard CNN models in a key aspect, namely, the max-pooling of the feature map layer of conventional CNN, which is employed in all modern CNN models for dimensionality reduction, is replaced by standard convolutional layers with stride two. You have to know the standard formula for defining convolutional and max-pooling operations in CNN if you want to understand why this procedure will work. Let ψ denote a feature map of input image attained by each convolutional layer of a CNN. It can be depicted as a three-dimensional array of size $W \times H \times G$, where H and W are the height and width and G is the number of channels. Then, p-norm max-pooling/subsampling with max-pooling size k ($k = 2$) and stride r is employed in the feature map ψ is a three-dimensional array $s(\psi)$ with the following entries [32]:

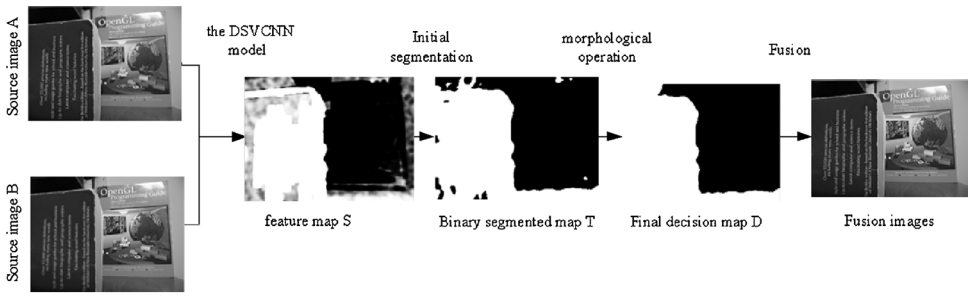
$$s_{i,j,u}(\psi) = \Big(\sum_{h=-\lfloor k/2 \rfloor}^{\lfloor k/2 \rfloor} \sum_{w=-\lfloor k/2 \rfloor}^{\lfloor k/2 \rfloor} |\psi_{g(h,w,i,j,u)}|^p\Big)^{1/p} \tag{8}$$

where $g(h, w, i, j, u) = (r{\cdot}i + h, r{\cdot}j + w, u)$ is the function mapped from locations in s to locations in ψ respecting the stride, and p is the order of the p-norm, which becomes the most commonly used max-pooling. When $r > k$, the max-pooling regions do not overlap; current emblematical CNN frameworks usually include overlapping max-pooling with $k = 3$ and $r = 2$. We compare the max-pooling operation to the standard definition of a convolutional layer c employed in feature map ψ, given as follows:

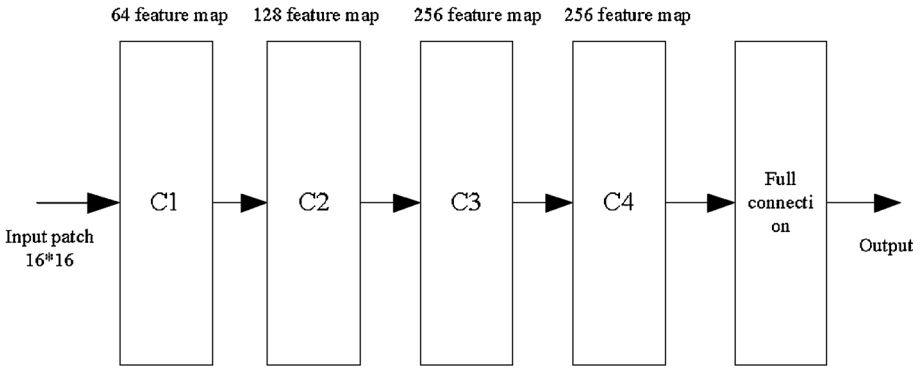
$$c_{i,j,o}(\psi) = f\Big(\sum_{h=-\lfloor k/2 \rfloor}^{\lfloor k/2 \rfloor} \sum_{w=-\lfloor k/2 \rfloor}^{\lfloor k/2 \rfloor} \sum_{u=1}^{N} \theta_{h,w,u,o} \cdot \psi_{g(h,w,i,j,u)}\Big) \tag{9}$$

where θ is the kernel weights (or the convolutional weights or filters), and i, j and u represent the coordinates of the pixels (i, j) at scale u. Here, $f(\bullet)$ is the activation function, which is usually a rectified activation ReLU, $f(x) = \max(x, 0)$, and $o \in [1, M]$ is the number of output characteristics of the convolutional layer. When this is formalized, it is obvious that these two operations rely on the same elements as the feature map in the previous layer. The max-pooling layer in the standard CNN model can be considered as performing a feature-wise convolution (a convolution where $\theta_{h,w,u,o} = 1$ if u equals o and zero otherwise), in which the activation function is replaced by the p-norm. The function of the max-pooling/subsampling layer has two main points: 1) dimensionality of the feature map; and 2) maintenance of the feature scale-invariant characteristics to a certain extent. It is easy to see that max-pooling can be removed from a convolutional neural network without relinquishing the spatial dimensionality reduction. The max-pooling of the feature map layer of conventional CNN, which is employed in all modern CNN models for dimensionality reduction, is replaced by standard convolutional layers with a stride of two (i.e., a max-pooling layer, in which $k = 3$ and $r = 2$ is replaced by a convolutional layer with a corresponding kernel and stride size).

The substitution of max-pooling by a convolutional layer increases the inter-feature dependencies only when the weight matrix θ is limited. We want to stress is that this substitution can be considered as learning the max-pooling operation, not repairing it; it has previously been used in the literature to consider the use of different parameterizations [33,34]. In multi-focus image fusion, although we do not know of existing research containing controlled experiments to replace max-pooling with a convolutional layer, it is worth pointing out that the study of eliminating max-pooling is not unprecedented. The naming used in early CNN work [33], which referred to max-pooling as a sub-sampling layer, indicates the use of different operations for subsampling. Although only small



(a) Schematic diagram of the proposed fusion algorithm



(b) The architecture of the DSVCNN

Fig. 3. Schematic diagram of the proposed fusion algorithm and the architecture of the CNN.

networks are considered, experiments that use only convolutions (occasional sampling) in an architecture similar to conventional CNN have appeared on the “neural abstraction pyramid” [35].

5. Method implementation

The schematic diagram of the proposed method is displayed in Fig. 3(a), which clearly shows that the proposed algorithm contains four steps: focus detection, initial segmentation, morphological operation and final fusion. In the first step, the two input images are provided to pre-train the proposed convolutional neural network model to output a feature map; the feature map contains the focus/clear information of source input images. Every coefficient in the feature map denotes the focus property of a pair of corresponding patches from the two source input multi-focus images. A focus/decision map with the same size as the source input images is produced from the feature map by averaging the overlapping patches. Step two, the feature map obtained from DSVCNN is segmented into a binary map with a fixed threshold. Step three; we optimize the binary segmented map with a mathematical morphological processing algorithm to produce the final decision map. In the final step, the fused image is produced by the ultimate decision map using the pixel-wise weighted-average strategy.

5.1. Focus detection

We assume that A and B respectively represent two original input images to be fused. In this study, if the source input image to be fused is a color image; it is first transformed into a grayscale image. Through the fusion method presented in this paper, we obtain the feature map S first, where the matrix S ranges from 0 to 1.

Fig. 3(a) shows that the focus information of the input image is detected accurately. It is generally observed that the value of the region with rich details is close to 0 (black) or 1 (white), while the plain region has its own value close to 0.5 (gray).

5.2. Decision map optimization

To obtain a satisfactory decision map for image fusion, feature map *S* must be further optimized. In the literature, the

representative and popular maximum strategy-based method is used to optimize the feature map S [36,2]. Correspondingly, a fixed threshold (0.65) is used to segment S into a binary segmented map T, The decision map T can be denoted as follows:

$$T(x, y) = \begin{cases} 1, & S(x, y) > 0.65 \\ 0, & otherwise \end{cases} \tag{10}$$

From Fig. 3(a), we can clearly see that the binary map T may contain many misclassified pixels and has some small holes, which can be easily removed by the mathematical morphological processing algorithm. Areas that are smaller than the region threshold are removed from the binary map. In this article, the area threshold is universally set to 0.01*Hei*Wid, where Hei and Wid are the height and width of input source image, respectively. To remove these defects, mathematical morphology methods are used in the following steps:

(1) First, use ceil to take an integer on the area.

$$\text{area} = \text{ceil}(0.01*Hei*Wid) \tag{11}$$

(2) Then, use the filter bwareaopen to remove the black or white area, as shown in the follow:

$$\text{Tm1} = \text{bwareaopen}(\text{T}, \text{area}) \tag{12}$$

$$\text{Tm2} = \text{bwareaopen}(1\text{-Tm1}, \text{area}) \tag{13}$$

$$D = 1 - \text{Tm2} \tag{14}$$

Fig. 3(a) shows the obtained final decision map D after applying mathematical morphological processing algorithm.

The fused image F is produced by considering the pixel weighted average rule from the final fusion decision map D as follows:

$$F(x, y) = D(x, y)A(x, y) + (1 - D(x, y))B(x, y) \tag{15}$$

5.3. Method implementation

The architecture of the DSVCNN is shown in Fig. 3(b). The computation of the proposed DSVCNN method can be summarized as follows.

(1) According to Eqs. (3)–(6), the central row vector of matrix Q is obtained; we reshape it into a weight kernel and then obtain the support value filter.
(2) The two source images are input to DSVCNN; the output image is the convolution operation of the input image and the support value filter.
(3) The first and second convolutional layer in the DSVCNN can obtain 64 feature maps and 128 feature maps using a 3×3 filter; the stride of the two convolutional layers is set to 1.
(4) The filter size in the third convolution is set to 3×3 and the stride layer is set to 2 to obtain 256 feature maps.
(5) These 256 feature maps are input to the fourth convolutional layer to obtain 256 feature maps using a 3×3 filter.
(6) The 256 feature maps are forwarded to the fully connected layer. The output of the DSVCNN is a two-dimensional vector.

A 2-way soft-max layer uses the 2-dimensional vector as input, and outputs two kinds of probability distributions [37].

Just as in CNN-based tasks [37], the soft-max function is employed in this study as the objective of the DSVCNN framework. The weight decay and the momentum are initialized to 0.0005 and 0.9 in our proposed DSVCNN training procedure. The weights and biases are updated layer by layer using the rules discussed in Section III. The proposed CNN framework uses the representative and popular learning framework in the literature to process the input image; we train the proposed model using Caffe in [38]. The parameters of every convolutional layer in the proposed DSVCNN model are initialized using the Xavier method. The leaning rate of each convolutional layer is identically set to 0.0001. Throughout the training process, the learning rate dropped once [26].

6. Experiments

Several pairs of input images are employed as test images in our experiments to examine the feasibility of the proposed DSVCNN-based fusion algorithm. We compared the multi-focus image fusion algorithm presented in this article with the recent state of the art multi-focus image fusion methods MWGF [25], SSDI [36], DCNN [26] and DSIFT [2]. A detailed discussion and analysis of the image fusion results is presented below.

6.1. Subjective evaluation of fused images

We compare the effectiveness of different multi-focus image fusion methods by considering visual quality first. To do this, the “Children” source image pair is used as an example to illustrate the fusion effect of different multi-focus image fusion methods. Fig. 4 shows the fused images of the “Children” source image obtained by the different compared image fusion methods. As shown in Fig. 4, the five algorithms can achieve the hoped-for goal of image fusion quality. However, different visual qualities of fused images are

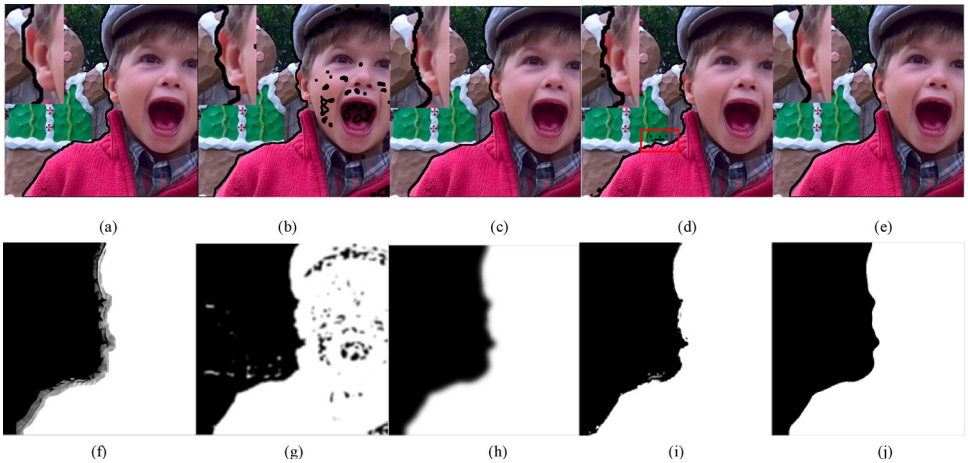


Fig. 4. The fused results of five methods on the 'Children' image set: Images (a)–(e) in the first row are the fusion results of, MWGF, SSDI, CNN,DSIFT and DSVCNN, respectively; each of the fused images clearly shows the multi-focus boundaries overlaid on the fusion image. Images (f)–(j) in the second row show the decision maps produced by MWGF, SSDI, CNN, DSIFT and DSVCNN, respectively.

produced by different multi-focus image fusion algorithms, according to their performance. To achieve a better comparison, the region around the boundary between the defocused and focused areas is clearly marked by a black line in each of the fused images (see the first row of Fig. 4). In Fig. 4 (b), the boundary line near the ear in the enlarged area is not smooth. The SSDI based algorithms produce some undesirable black spots or lines in the fused image, indicating that the children's faces also have some multi-focus boundaries, but this is impossible. This result directly reveals the shortcomings of the SSDI based method in multi-focus image fusion.

Similar to the SSDI-based fusion methods, the DSIFT-based methods also produce an incorrect boundary classification between the defocused and focused areas in the rectangular area (see Fig. 4(d)). The algorithm based on DSIFT often cannot achieve a satisfactory image fusion quality from the source image. To clearly display the details of the fused results, in each of the fused images, the partial regions around the boundary between the defocused and focused areas are zoomed and shown in the upper left corner.

From the zoomed regions of Fig. 4, it can be seen that there are many jagged phenomena in the boundary area apart from Fig. 4(e). Fig. 4(e) shows that the fused image of the DSVCNN-based method is quite satisfactory, and the boundary shown in Fig. 4(e) is relatively smooth compared to the other image fusion algorithms. Finally, because of the superiority of the algorithms presented in this article, DSVCNN accurately detects the boundary of the input multi-focused image between the defocused and focused regions and then produces a better decision map from the source input images than do the other four image fusion methods in this study. The fusion result of the DSVCNN based method displays the best subjective evaluation index compared to the other four algorithms.

The fusion images are difficult to evaluate thoroughly using only the human visual system. For a thorough test of the feasibility of the DSVCNN-based method for multi-focus image fusion, we compare the decision maps obtained by the five image fusion methods. From the decision maps (as shown in Figs. 5–7), the advantages and disadvantages of the five images fusion algorithms can be seen clearly. Comparisons of the decision maps for the "Children" image are shown in the second row of Fig. 4. From Fig. 4(f), we can see

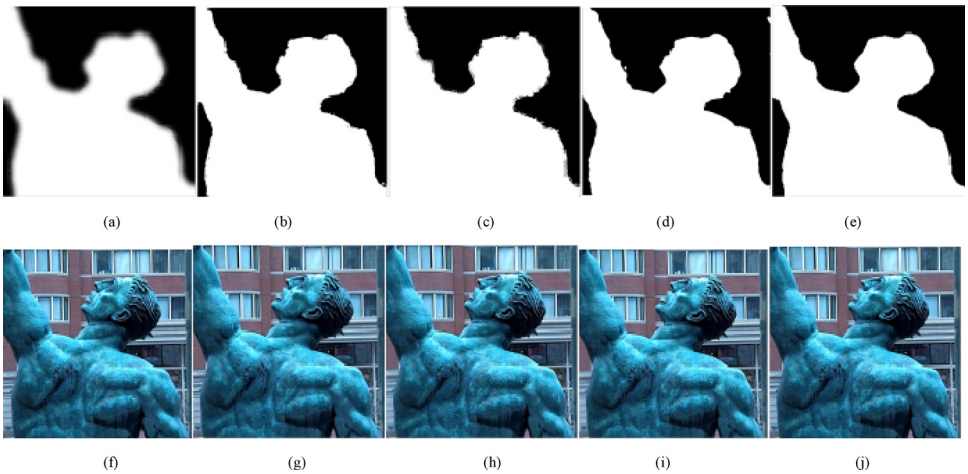


Fig. 5. The first row shows the decision maps produced by CNN, DSIFT, MWGF, SSDI and DSVCNN; the second row shows the fusion results of CNN, DSIFT, MWGF, SSDI and DSVCNN on the 'Man' image set.

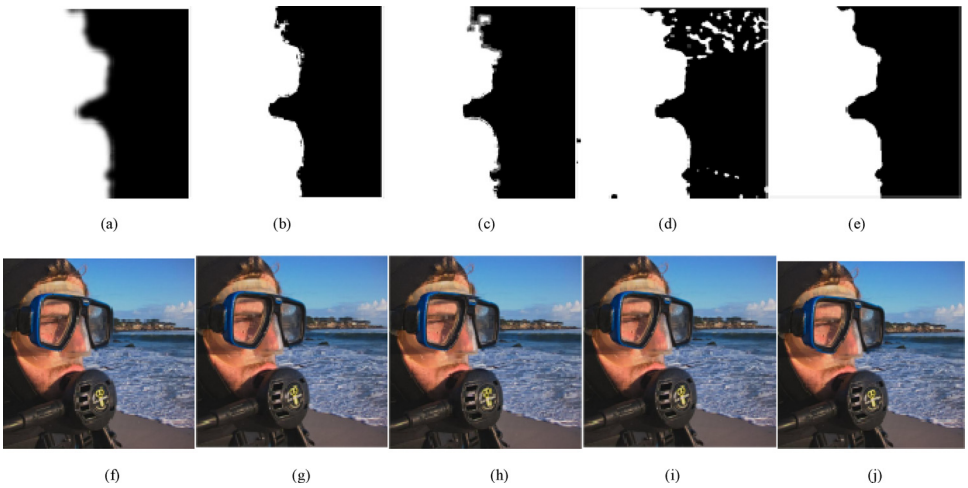


Fig. 6. The first row shows the decision maps produced by CNN, DSIFT, MWGF, SSDI and DSVCNN; the second row shows the fusion results of CNN, DSIFT, MWGF, SSDI and DSVCNN on the 'Diver' image set.

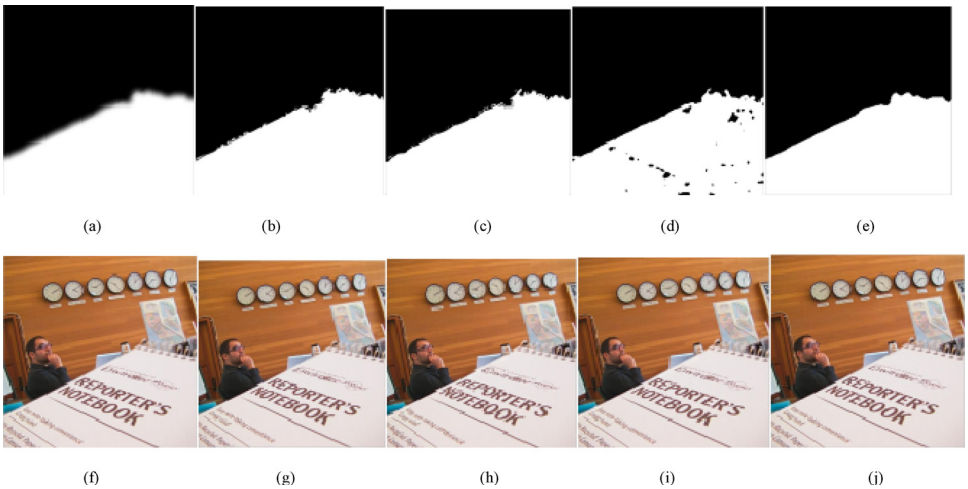


Fig. 7. The first row shows the decision map obtained by CNN, DSIFT, MWGF, SSDI and DSVCNN; the second row shows the fusion results of CNN, DSIFT, MWGF, SSDI and DSVCNN on the 'Notebook' image set.

that the decision map obtained by the MWGF-based method has obvious shadows in the boundary area. The right side of the decision map shows some black spots (see Fig. 4(g)), which means that there are weaknesses in the image in the fusion method based on SSDI. In Fig. 4(i), the decision map produced by the DSIFT-based algorithm is jagged near the boundary area. Fig. 4(h) and (j) show that the fused images of the DSVCNN and CNN-based methods are quite satisfactory, and the boundaries displayed in Fig. 4(h) and Fig. 4(j) are relatively smooth compared with those of the other methods. However, with respect to Fig. 4(h), the contour of the boundary of the decision map in Fig. 4(j) is closer to the children.

To further illustrate the effectiveness of the DSVCNN method for multi-focus image fusion, Figs. 5–7 show example decision maps and fusion images produced by the five multi-focus image fusion methods. In these decision maps, the pros and cons of the various fusion methods are clearly visible. The "choose-max" strategy method is used in the binary segmentation algorithm of the proposed image fusion algorithm to produce a binary segmented decision map from the feature map (see Fig. 3(a)) with a fixed threshold. For the multi-focus image fusion problem, the binary feature map in Fig. 3(a) is the actual output of our DSVCNN-based method. From the binary segmented map in Fig. 3(a), we can conclude that the segmented maps produced by the DSVCNN-based method are highly effective because the great majority of pixels are accurately classified, which indicates the success of the DSVCNN-based method.

In this study, we mainly consider the situation that there are only two pre-registered source images. To deal with more than two multi-focus images, one can fuse them one by one in series. To prove that the proposed method can be extended to multi-focus set containing more than two images, the fused results of two groups of the triple series are shown in Fig. 8.

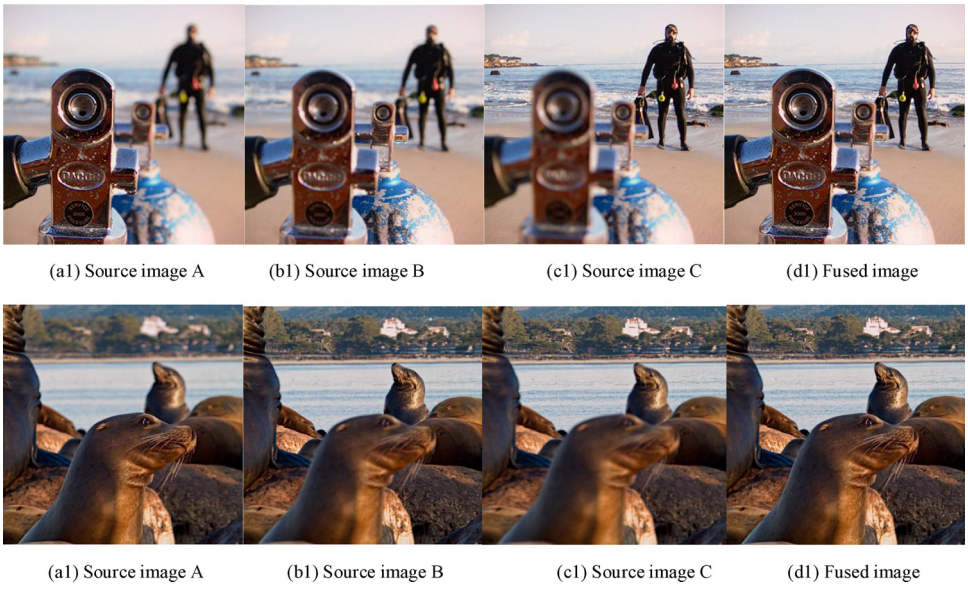


Fig. 8. The fused results of two groups of the triple series.

6.2. A fair comparison

The comparison in Figs. 4, 6 and 7 is not fair for MWGF, SSDI and DSIFT as a result of the proposed method and CNN adopts morphological filter to eliminate holes in the initial decision map. In this subsection, this morphological filter also applied to other methods to achieve a fair comparison. Fig. 9 shows the decision map obtained by using morphological filters for DSIFT, MWGF and SSDI. From Fig. 9, one can clearly see the advantages and disadvantages of various fusion methods.

6.3. Objective evaluation of fused images

For objective evaluation of fused results, two conventional indexes of mutual information, MI and $Q^{AB/F}$, are used as the quantitative evaluation criteria of image fusion performance (see Table 1) [39–42]. The quantitative evaluation criteria MI and $Q^{AB/F}$ are calculated on the fused images from the five fusion methods in this article and listed in Table 1. We can conclude that the method based on DSVCNN provides the best fusion results by considering the metrics MI except for the "Man" image. Based on the $Q^{AB/F}$ index scores, the DSVCNN-based method yields satisfactory fusion results for the source images of the "Note work", "Lab", "Book" and "Leopard" images, while the DSIFT method outperforms the DSVCNN-based method for the "Man", "Temple" and "Seascape" images, and the CNN-based method outperforms the DSVCNN-based method for the test images "Children" and "Flower". These results demonstrate that the DSVCNN-based fusion method needs further improvement and development to protect the edge information of the input image during the fusion process: the $Q^{AB/F}$ index considers a fused image that contains all the input image edge information as the ideal fusion result.

In addition to the above two commonly used objective criteria, two novel objective criteria Q_Y and Q_P, which were used in [2], are employed in this article to evaluate various fusion methods (see Table 2). The quantitative evaluation criteria Q_Y and Q_P of the fused images using the five fusion methods in this article are listed in Table 2, from which we can see that the method based on the DSVCNN proposed in this article provides the best fusion results when considering the metrics Q_Y except for the "Man" image. From the Q_P index scores, we can conclude that the DSVCNN-based method achieves satisfactory fusion results for the source images except on the "Note work", "Man", "Temple" and "Diver", "Seascape" and "Book" images in Table 2. Q_Y is an image fusion metric based on structural similarity, which measures the level of structural information of the source images preserved in the fused image. From Table 2, one can conclude that the DSVCNN-based fusion method better protects the structural information of source images. Because Q_P is a feature-based objective assessment, the results in the table demonstrate that the DSVCNN-based fusion method needs further improvement to protect the feature information of input images in the fusion process.

Table 3is the objective assessments obtained by using morphological filters for DSIFT, MWGF and SSDI. After using morphological filtering for DSIFT, MWGF and SSDI, the objective evaluation scores of the fused image is improved, and even the value of MI on 'Children' image set is more than the proposed method. In general, even though the other four methods all use morphological filtering, the objective evaluation scores of the proposed method are the best in the vast number of cases. However, compared with several other algorithms, the proposed method is very time consuming.

Fig. 9. The decision map obtained by using morphological filters for DSIFT, MWGF and SSDI. The first, second, third row and last row show the decision maps produced by DSIFT, MWGF, SSDI and proposed method, respectively. The first column, the second column and last column show the decision maps on the 'Children', 'Diver' and 'Note work' image sets.

Table 1
Comparison of quantitative evaluation criteria: the MI and $Q^{AB/F}$ scores of the tested methods.

		MWGF	SSDI	CNN	DSIFT	DSVCNN
Lab	MI	8.0618	8.1412	8.6008	8.5201	**8.8333**
	$Q^{AB/F}$	0.7147	0.7528	0.7573	0.7585	**0.7587**
Man	MI	8.1901	8.3077	8.4138	**8.7729**	8.7636
	$Q^{AB/F}$	0.7689	0.7108	0.7815	**0.7789**	0.7787
Temple	MI	5.9655	7.0896	6.8895	7.3514	7.4015
	$Q^{AB/F}$	0.7501	0.7634	0.7590	**0.7643**	0.7642
Diver	MI	8.8766	8.6595	9.1534	9.3090	**9.3330**
	$Q^{AB/F}$	0.7524	0.7103	**0.7560**	0.7550	0.7549
Seascape	MI	7.1404	7.4824	7.6285	7.9487	**8.0077**
	$Q^{AB/F}$	0.7059	0.7110	0.7113	0.7126	**0.7132**
Note work	MI	8.2413	8.4785	8.5259	8.7389	**8.7852**
	$Q^{AB/F}$	0.7766	0.7010	**0.7811**	0.7810	**0.7799**
Book	MI	8.2368	8.4008	8.7796	8.6623	**8.8747**
	$Q^{AB/F}$	0.7240	0.7260	0.7277	0.7134	**0.7281**
Leopard	MI	9.9474	10.8887	10.8792	10.9226	**10.9392**
	$Q^{AB/F}$	0.8175	0.8171	0.7973	0.8069	**0.8271**
Children	MI	8.2622	7.8505	8.3338	8.5252	**8.5401**
	$Q^{AB/F}$	0.6741	0.6799	0.7408	**0.7394**	0.7393
Flower	MI	8.3255	8.1049	8.2659	8.5365	**8.5818**
	$Q^{AB/F}$	0.6913	0.6490	**0.7183**	0.7159	0.7161

Table 2
Comparison of objective assessments: the Q_Y and Q_P scores of the tested methods.

		MWGF	SSDI	CNN	DSIFT	DSVCNN
Lab	Q_Y	0.9724	0.8980	0.9780	0.9654	**0.9892**
	Q_P	0.7986	0.7860	0.8047	0.7958	**0.8053**
Man	Q_Y	0.9716	0.9775	**0.9783**	0.9647	0.9777
	Q_P	0.8963	0.9081	**0.9804**	0.9182	0.9063
Temple	Q_Y	0.9897	0.9904	0.9927	0.9913	**0.9945**
	Q_P	0.7771	**0.7917**	0.7889	0.7816	0.7832
Diver	Q_Y	0.9809	0.9831	0.9881	0.9851	**0.9888**
	Q_P	0.8936	**0.9017**	0.8985	0.8982	0.8975
Seascape	Q_Y	0.9874	0.9572	0.9932	0.9824	**0.9952**
	Q_P	0.6659	**0.6849**	0.6729	0.6548	0.6703
Note work	Q_Y	0.9860	0.9913	0.9948	0.9902	**0.9961**
	Q_P	0.7914	**0.8018**	0.7965	0.7934	0.7921
Book	Q_Y	0.9702	0.9715	0.9780	0.9726	**0.9892**
	Q_P	0.8658	**0.9325**	0.8047	0.8012	0.8052
Leopard	Q_Y	0.9860	0.9932	0.9904	0.9889	**0.9933**
	Q_P	0.7914	0.9330	0.9414	0.9452	**0.9514**
Children	Q_Y	0.9072	0.9458	0.9868	0.9815	**0.9910**
	Q_P	0.8572	0.8626	0.8619	0.8521	**0.8746**
Flower	Q_Y	0.9774	0.9802	0.9830	0.9800	**0.9831**
	Q_P	0.7885	0.7968	0.7944	0.7864	**0.7991**

Table 3
Objective assessments obtained by using morphological filters for DSIFT, MWGF and SSDI.

		MWGF	SSDI	CNN	DSIFT	DSVCNN
Children	MI	8.5297	8.5386	8.3338	**8.5443**	8.5401
	$Q^{AB/F}$	0.7296	0.7378	**0.7408**	0.7398	0.7393
	Q_Y	0.9145	0.9548	0.9868	0.9836	**0.9910**
	Q_P	0.8462	0.8550	0.8619	0.8578	**0.8746**
Diver	MI	9.2798	9.2157	9.1534	9.3159	**9.3330**
	$Q^{AB/F}$	0.7515	0.7500	**0.7560**	0.7551	0.7549
	Q_Y	0.9884	0.9812	0.9881	0.9885	**0.9888**
	Q_P	0.8945	0.8912	**0.8985**	0.8970	0.8975
Note work	MI	8.7908	8.7678	8.5259	8.7483	**8.7852**
	$Q^{AB/F}$	0.7758	0.7768	**0.7811**	0.7815	**0.7799**
	Q_Y	0.9956	0.9935	0.9948	0.9958	**0.9961**
	Q_P	0.7830	0.7891	**0.7965**	0.7944	0.7921
average time (s)		4.500	39.00	238.2	11.00	254.0

7. Conclusions

In this paper, we presented a novel multi-focus fusion method based on deep support values to address the shortcomings of fusion methods based on the CNN. A new fusion framework based on DSVCNN is presented. The experimental results demonstrate the advantages of the proposed DSVCNN-based methods over CNN-based methods. This is the first time that DSVCNN has been applied to multi-focus image fusion and the second time that CNN has been employed for multi-focus image fusion. As evidenced by the great progress achieved in the CNN-based multi-focus image fusion algorithm, more efficacious multi-focus image fusion CNN-based methods can be developed and applied in the field of image fusion to pursue better fusion performance. We believe that DSVCNN can be the start of a new research approach to the field of multi-focus image fusion.

Acknowledgments

The work of this paper was supported by the National Natural Science Foundation of China (Project Number: 61174193) and the Specialized Research Fund for the Doctoral Program of Higher Education (Project Number: 20136102110036).

References

[1] S. Li, B. Yang, Hybrid multiresolution method for multisensor multimodal image fusion, IEEE Sens. J. 10 (September (9)) (2010) 1519–1526.
[2] Y. Liu, S. Liu, Z. Wang, Multi-focus image fusion with dense SIFT, Inf. Fusion 23 (May) (2015) 139–155.
[3] Q. Zhang, B.L. Guo, Multifocus image fusion using the nonsubsampled contourlet transform, Signal. Process. 89 (July) (2009) 1334–1346.
[4] S. Li, B. Yang, J. Hu, Performance comparison of different multiresolution transforms for image fusion, Inf. Fusion 12 (April) (2011) 74–84.
[5] C. Du, S. Gao, Image segmentation-based multi-focus image fusion through multi-scale convolutional neural network, IEEE Access. 5 (2017) 15750–15761.
[6] V.N. Gangapure, S. Banerjee, A.S. Chowdhury, Steerable local frequency based multispectral multifocus image fusion, Inf. Fusion 23 (May) (2015) 99–115.
[7] S. Pertuz, D. Puig, M.A. Garcia, A. Fusiello, Generation of all-in focus images by noise-robust selective fusion of limited depth-of-field images, IEEE Trans. Image Process. 22 (March (3)) (2013) 1242–1251.
[8] L. Cao, L. Jin, H. Tao, G. Li, Z. Zhuang, Y. Zhang, Multi-focus image fusion based on spatial frequency in discrete cosine transform domain, IEEE Signal. Process. Lett. 22 (February (2)) (2015) 220–224.
[9] Y.P. Liu, J. Jin, Q. Wang, Y. Shen, X. Dong, Region level based multi-focus image fusion using quaternion wavelet and normalized cut, Signal. Process. 97 (April) (2014) 9–30.
[10] L. Guo, M. Dai, M. Zhu, Multifocus color image fusion based on quaternion curvelet transform, Opt. Exp. 20 (17) (2012) 18846–18860.
[11] Q.G. Miao, C. Shi, P.F. Xu, M. Yang, Y.B. Shi, A novel algorithm of image fusion using shearlets, Opt. Commun. 284 (6) (2011) 1540–1547.
[12] Y. Chai, H. Li, X. Zhang, Multifocus image fusion based on features contrast of multiscale products in nonsubsampled contourlet transform domain, Optik-Int. J. Light Electron. Opt. 123 (April) (2012) 569–581.
[13] S. Li, X. Kang, L. Fang, J. Hu, H. Yin, Pixel-level image fusion: a survey of the state of the art, Inf. Fusion 33 (2017) 100–112.
[14] A. Goshtasby, S. Nikolov, Image fusion: advances in the state of the art, Inf. Fusion 8 (2) (2007) 114–118.
[15] S. Li, X. Kang, J. Hu, B. Y, Image matting for fusion of multi-focus images in dynamic scenes, Inf. Fusion 14 (2) (2013) 147–162.
[16] W. Zhang, W.K. Cham, Gradient-directed multi-exposure composition, IEEE Trans. Image Process. 21 (4) (2012) 2318–2323.
[17] B. Gu, W. Li, J. Wong, M. Zhu, M. Wang, Gradient field multi-exposure images fusion for high dynamic range image visualization, J. Vis. Commun. Image Represent. 23 (4) (2012) 604–610.
[18] S. Li, J. Kwok, Y. Wang, Combination of imageswith diverse focuses using the spatial frequency, Inf. Fusion 2 (3) (2001) 169–176.
[19] A. Goshtasby, Fusion of multi-exposure images, Image Vis. Comput. 23 (6) (2005) 611–618.
[20] V. Aslantas, R. Kurban, Fusion of multi-focus images using differential evolution algorithm, Expert Syst. Appl. 37 (12) (2010) 8861–8870.
[21] X. Bai, Y. Zhang, F. Zhou, B. Xue, Quadtree-based multi-focus image fusion using a weighted focus-measure, Inf. Fusion 22 (1) (2015) 105–118.
[22] M. Li, W. Cai, Z. Tan, A region-based multi-sensor image fusion scheme using pulse-coupled neural network, Pattern Recognit. Lett. 27 (16) (2006) 1948–1956.
[23] S. Li, X. Kang, J. Hu, Image fusion with guided filtering, IEEE Trans. Image Process. 22 (July (7)) (2013) 2864–2875.
[24] H. Li, Y. Chai, H. Yin, G. Liu, Multifocus image fusion and denoising scheme based on homogeneity similarity, Opt. Commun. 285 (2) (2012) 91–100.
[25] Z. Zhou, S. Li, B. Wang, Multi-scale weighted gradient-based fusion for multi–focus images, Inf. Fusion 20 (2014) 60–72.
[26] Y. Liu, X. Chen, H. Peng, Z. Wang, Multi-focus image fusion with a deep convolutional neural network, Information Fusion 36 (2017) 191–207.
[27] S. Guo, S. Chen, Y. Li, Face recognition based on convolutional neural network & support vector machine, IEEE ICIA 2016 (2017) 1787–1792 January 24.
[28] Y. Cao, R. Xu, T. Chen, Combining convolutional neural network and support vector machine for sentiment classification, Commun. Comput. Inform. Sci. 568 (2015) 144–155.
[29] S. Zheng, W. Shi, J. Liu, et al., Multisource image fusion method using support value transform, IEEE Trans. Image Process. 16 (7) (2007) 1831–1839.
[30] S. Zheng, W. Shi, J. Liu, J.W. Tian, Remote sensing image fusion using multiscale mapped LS-SVM, IEEE Trans. Geosci. Remote Sens. 46 (5) (2008) 1313–1322.
[31] Y. LI, Z. HAO, H. LEI, Survey of convolutional neural network, J. Comput. Appl. 36 (9) (2016) 2508–2515 256.
[32] Jost Tobias Springenberg, Alexey Dosovitskiy, Striving for Simplicity: The All Convolutional Net. In ICLR, (2015).
[33] Y. LeCun, L. Bottou, Y. Bengio, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324.
[34] Y. Jia, C. Huang, Darrell, Beyond Spatial Pyramids: Receptive Field Learning for Pooled Image Features. In CVPR, (2012).
[35] Sven Behnke, Hierarchical Neural Networks for Image Interpretation. PhD Thesis, (2003).
[36] D. Guo, J.W. Yan, X. Qu, High quality multi-focus image fusion using self-similarity and depth information, Opt. Commun. 338 (2015) 138–144.
[37] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic seg- mentation, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (2015), pp. 3431–3440.
[38] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadar- rama, T. Darrell, Caffe: convolutional architecture for fast feature embedding, Proceedings of the ACM International Conference on Multimedia, (2014), pp. 675–678.
[39] B. Yang, S. Li, Pixel-level image fusion with simultaneous orthogonal matching pursuit, Inf. Fusion 13 (2012) 10–19.
[40] Chaoben Du, Shesheng Gao, Multi-focus image fusion algorithm based on pulse coupled neural networks and modified decision map, Optik 157 (2018) 1003–1015.
[41] G. Piella, H. Heijmans, A new quality metric for image fusion, Proc. IEEE Int. Conf. Image Process (2003) 173–176.
[42] G. Bhatnagar, Q. M, Directive contrast based multimodal medical image fusion in NSCT domain, IEEE Trans. Multimedia 15 (5) (2013) 1014–1024.