

Received July 5, 2017, accepted July 29, 2017, date of publication August 2, 2017, date of current version August 29, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2735019

Image Segmentation-Based Multi-Focus Image Fusion Through Multi-Scale Convolutional Neural Network

CHAOBEN DU AND SHESHENG GAO

School of Automation, Northwestern Polytechnical University, Xi'an 710129, China

Corresponding author: Chaoben Du (dcbxjdaxue@163.com)

This work was supported in part by the National Natural Science Foundation of China under Project 61174193 and in part by the Specialized Research Fund for the Doctoral Program of Higher Education under Project 20136102110036.

ABSTRACT A decision map contains complete and clear information about the image to be fused, and detecting the decision map is crucial to various image fusion issues, especially multi-focus image fusion. Nevertheless, in an attempt to obtain an approving image fusion effect, it is necessary and always difficult to obtain a decision map. In this paper, we address this problem with a novel image segmentation-based multi-focus image fusion algorithm, in which the task of detecting the decision map is treated as image segmentation between the focused and defocused regions in the source images. The proposed method achieves segmentation through a multi-scale convolutional neural network, which performs a multi-scale analysis on each input image to derive the respective feature maps on the region boundaries between the focused and defocused regions. The feature maps are then inter-fused to produce a fused feature map. Afterward, the fused map is post-processed using initial segmentation, morphological operation, and watershed to obtain the segmentation map/decision map. We illustrate that the decision map gained from the multi-scale convolutional neural network is trustworthy and that it can lead to high-quality fusion results. Experimental results evidently validate that the proposed algorithm can achieve an optimum fusion performance in light of both qualitative and quantitative evaluations.

INDEX TERMS Convolutional neural network, multi-focus image, decision map, image fusion.

I. INTRODUCTION

In digital imaging, the imaging equipment usually has difficulty in shooting the target image, in which all of the objects of the image are effectively captured in focus. Normally, by setting an affirmative focal length for the optical lens, only the objects in the depth of field (DOF) are clear in the picture, while other objects can be indistinct. Fortunately, multi-focus image fusion technology has emerged to address the above-mentioned problems by integrating significant sharp functions from multiple images of the same scene. Over the past several years, a variety of image fusion algorithms have emerged. These fusion algorithms can be divided into two categories [1]: spatial domain algorithms and transform domain algorithms.

The image fusion algorithm based on the transform domain usually converts the source image to another feature domain, where the source image can be effectively fused. The most popular transform domain fusion algorithms are founded

on multi-scale transform (MST) methods. Some representative examples include the Laplacian pyramid (LP) [2], the morphological pyramid (MP) [3], the discrete wavelet transform (DWT) [4], the dual-tree complex wavelet transform (DTCWT) [5] and the non-subsampled contourlet transform (NSCT) [6]. These methods must go through three steps to fuse the image in terms of, the decomposition, fusion and reconstruction [7]. Many studies have also been conducted while taking this approach [8]–[9], where the input image is first transformed into a multi-resolution representation by multi-resolution. Then, they select the different spectral information and combine it to reconstruct the fused images.

A new transform domain fusion approach [10]–[14] has become a compelling branch of the field. Unlike the MST-based approach described above, these fusion algorithms transform the image into a single scale feature area through some superior signal theories, for example, Sparse Representation (SR) and Independent Component

Analysis (ICA). This type of approach usually uses the sliding window method approach after the approximate translation invariant fusion process. The most important problem with these approaches is in exploring a valid feature domain to obtain the focus map.

The block-based image fusion technique decomposes the input images into blocks, for example, an interesting block-based method based on pulse coupled neural network (PCNN) is presented in [15]; then, every pair in the block is fused first through a designed activity level measurement such as sum-modified-Laplacian(SML) [16]. Obviously, the size of the block has a large impact on the fusion results. Currently, many improved algorithms have emerged to replace the use of an artificial fixed block size in the previous block-based algorithms [17], [18]. For example, there are new adaptive block-based methods [19] using a differential evolution algorithm to obtain an optimum block size. Using the recently introduced method based on the quad-tree [20], [21], the input images can be adaptively split into blocks of different sizes according to the information in the image itself. Some spatial domain-based fusion algorithms are founded on image segmentation and the impact of the segmentation accuracy on the fusion quality is critical [22], [23].

With regard to both the transform domain and spatial domain image fusion methods, the fusion map is the crucial factor. To further enhance the quality of the image fusion, many of the recently proposed methods have become increasingly complex. In recent years, the multi-focus image fusion algorithm based on the spatial domain has been widely discussed. The simplest pixel-based image fusion algorithm directly averages the pixel values of all of the source images. The advantages of the direct averaging method are simple and fast, but its fused images tend to produce blurring effects, thus losing some of the original image information. To overcome the shortcomings of the direct average algorithm, several state-of-the-art pixel-based image fusion algorithms have been proposed, such as guided filtering [24] and dense SIFT [25]. Guided filtering and dense SIFT first generate the fusion map by detecting the focused pixels from each source image; then, based on the modified decision map, the final fused image is obtained by selecting the pixels in the focus areas. Decision map is the focus region detection map, in which, the white region indicates the focus region of Source A, whereas the black region indicates the focus region of Source B. Using the detected focused regions as a fusion decision map to guide the multi-focus image fusion process not only increases the robustness and reliability of the fusion results, but also reduces the complexity of the procedure. The multi-scale weighted gradient method is to reconstruct I_F by making its gradient as close as possible to $\nabla \hat{I}$, rather than according to the decision map [26]. Although these new algorithms can improve the visual quality of the fused images, they can lose some of the original image information due to inaccurate fusion decision maps.

The multi-focus image fusion can be treated as an image segmentation (binary classification) problem have been proposed in the literatures [18], that is, the generation of decision map in multi-focus image fusion can be treated as a binary segmentation problem. Specifically, the role of multi-focus image fusion rule is analogous to that of segmentation used in general image segmentation tasks. Thus, it is feasible to use CNN for image fusion in theory.

In this paper, we introduce convolutional neural networks. Although convolution neural networks have been successfully applied in the field of face recognition, license plate recognition [27], behaviour recognition [28], speech recognition [29] and image classification [30], there are few applications for image fusion work. We solved the problem mentioned above with a novel image segmentation-based image fusion method, in which the task of detecting the decision map is treated as image segmentation between the focused and defocused regions in the source images. The proposed method achieves segmentation through a multi-scale convolutional neural network (MSCNN), which conducts multi-scale analysis on each input image to derive the individual feature maps on the region boundaries between the focused and defocused regions. Feature maps are then inter-fused to produce a fused feature map. Additionally, the fused map is post-processed using initial segmentation, morphological operation and the watershed transform to obtain the segmentation map/decision map. We illustrate that the decision map obtained from the MSCNN is trustworthy and that it can lead to high-quality fusion results. Experimental results evidently validate that the proposed algorithm can obtain optimum fusion performance in the light of both qualitative and quantitative evaluations.

The remainder of this paper is organized as follows. The related theory of the convolutional neural network (CNN) method is introduced in Section 2. In Section 3, the proposed MSCNN-based fusion method is discussed in detail. In Section 4, the detailed results and discussions of the experiments are presented. Finally, in Section 5, we conclude the paper.

II. CONVOLUTIONAL NEURAL NETWORK

CNN is an emblematical depth learning model that attempts to learn a hierarchical representation of an image at different abstraction levels [31]. As shown in Fig. 1, a typical CNN is mainly composed of an input layer, convolution layer, max-pooling(subsampling), fully connected layer and output layer.

The input of the convolution neural network is usually the original image X . In this paper, we use H_i to represent the feature map of the i -th layer of the convolution neural network ($H_0 = X$). Assuming that H_i is the convolution layer, the generation process of H_i can be described as follows:

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

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

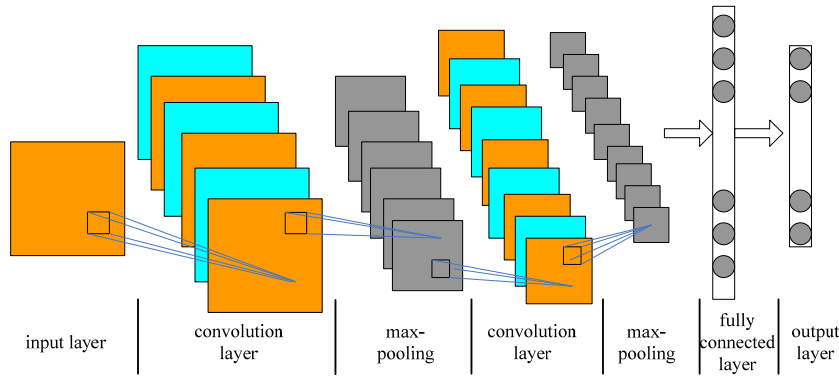


FIGURE 1. Typical Structure of a Convolution Neural Network.

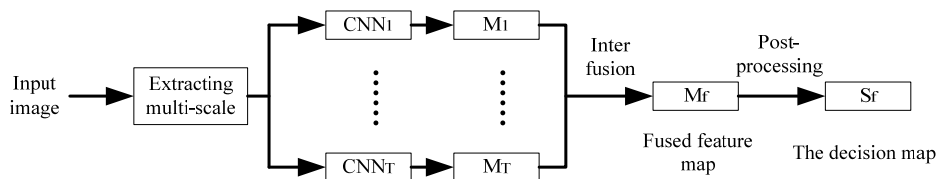


FIGURE 2. The conceptual work flow of the proposed method: T is the total number of CNN, Mc is a feature map of the location of boundary pixels of index c (c = 1,...,T), Mf is the fused feature map, and Sf is the decision map.

The max-pooling/subsampling layer usually follows the convolution layer; then, there is max-pooling of the feature graph according to a certain max-pooling rule. Through the alternation of multiple convolution and max-pooling layers, the convolution neural network relies on a fully connected network to classify the extracted features to obtain the probability distribution Y based on the input. The convolution neural network is essentially a mathematical model that makes the original matrix H_0 pass through multiple levels of data transformation or dimension reduction, mapping to a new feature expression Y .

$$Y(i) = P(L = l_i | H_0; (W, b)) \tag{2}$$

The training objective of the CNN is to minimize the loss function $L(W, b)$ of the network. The residuals are propagated backward by gradient descent, and the training parameters (W and b) of the individual layers of the convolution neural network are updated layer by layer.

$$W_i = W_i - \eta \frac{\partial E(W, b)}{\partial W_i} \tag{3}$$

$$b_i = b_i - \eta \frac{\partial E(W, b)}{\partial b_i} \tag{4}$$

where, $E(W, b) = L(W, b) + \frac{\lambda}{2} W^T W$, λ is the parameter of weight decay, and η is the learning rate parameter used to control the intensity of the residual back propagation.

As described above, the generation of decision map in multi-focus image fusion can be treated as a binary segmentation problem. For a pair of image patches $\{p_A, p_B\}$ of the same scene, our goal is to learn an CNN whose output is a

scalar ranging from 0 to 1. Specifically, when p_A is focused while p_B is the defocused region, the output value should be close to 1, and when p_B is focused while p_A is the defocused region, the output value should be close to 0. That is to say, the output value represents the focus property of the patch pair. Therefore, the use of the CNN to fuse the image in theory is feasible.

III. MULTI-SCALE CONVOLUTIONAL NEURAL NETWORK

A. METHOD FORMULATION

The conceptual work flow of the proposed MSCNN method is demonstrated in Fig. 2. The schematic diagram of the proposed multi-focus image fusion algorithm is shown in Fig. 3. From Fig. 3, we can see that the proposed multi-focus image fusion algorithm consists of four steps: MSCNN, initial segmentation, morphological operation and watershed, and the last step, fusion. In the first step, the two source images are fed into a pretreatment training MSCNN model to produce a feature map, and this map, includes the most focused information from the source images. Notably, the coefficient in the map represents the focus property of the patch that corresponds to the two source images [37]. Through averaging the overlapping patches, we obtain the feature map of the focus map with the same size of the source image in this paper. In the second step, the feature map is split into a binary map with a threshold of 0.9. In the third step, we extract and smoothen the binary segmented map with the morphological operation and watershed to generate the final decision map (The filter `bwareaopen` is employed to remove the black area as Morphological operations, and the

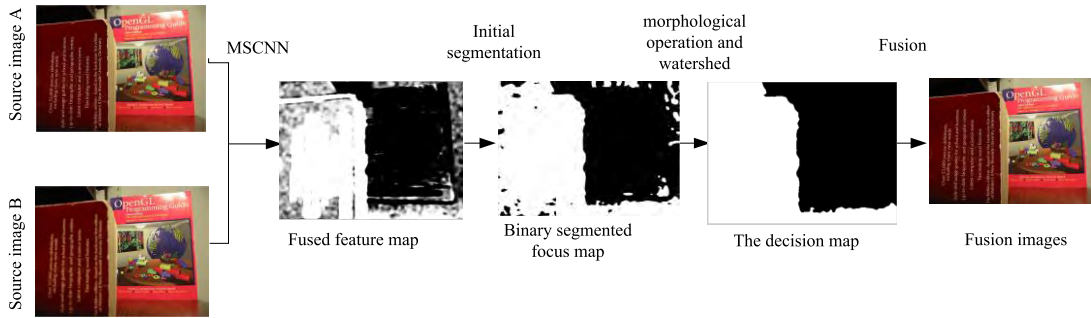


FIGURE 3. Schematic diagram of the proposed multi-focus image fusion algorithm.

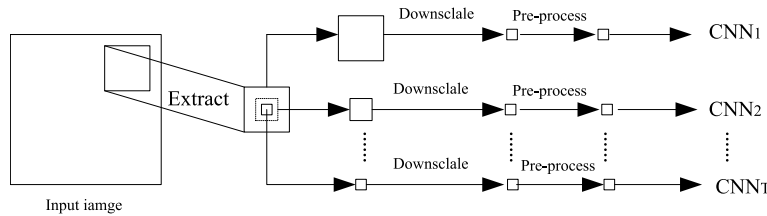


FIGURE 4. The process of extracting multi-scale input patches from the input image.

threshold of the bwareaopen depends on the image size, that is, $0.01 * I_H * I_W$, where I_H and I_W are the length and width of the image, respectively). In the last step, the fused image is obtained in the final decision map by the pixel-wise weighted-average strategy.

B. DECISION MAP OPTIMIZATION

We assume that A and B represent two original images that are going to be fused. In this paper, if the image to be fused is a colour image, then we transform it into the grey space first. Through the method proposed in this paper, we obtain the feature map M_f first, and the matrix S range is from 0 to 1. As seen from the feature map of Fig. 3, the most focused information of the source image is accurately detected. Directly perceived through the senses, the value of the area with rich details appears to be close to 1 or 0, while the plain area tends to have its own value close to 0.5.

To obtain a more accurate decision map, the feature map M_f must be further processed in this algorithm. The most popular maximum strategy is used to process the feature map M_f [38], [39]. Correspondingly, we use a fixed threshold $\beta = 0.9$ to segment S into binary segmented map D. The map D is given as follows:

$$D(x, y) = \begin{cases} 1, & S(x, y) > 0.9 \\ 0, & otherwise \end{cases} \quad (5)$$

From Fig.2, it can be seen that the binary map Bm could contain some misclassified pixels, and these error categories can be easily removed through the small area clear strategy. The fused feature map sometimes contains some very small holes. When these holes occur, we should also use morphological processing.

Combined with the final fusion decision map S_f , the fused image F is obtained according to the pixel weighted average

rule, as follows:

$$F(x, y) = S_f(x, y)A(x, y) + (1 - S_f(x, y))B(x, y) \quad (6)$$

C. MULTI-SCALE ANALYSIS

Assume that the input image is I, and let $I = \{p(x, y) : 1 \leq x \leq X, 1 \leq y \leq Y\}$, where $p(x, y)$ is the value of a pixel at (x, y) in the input image, I, with $X \times Y$ resolution. Assume that a patch $P(x, y)$ is the $w \times w$ window that surrounds the pixel (x, y) in the input image, which is defined as

$$P(x, y) = \{p(x - \lfloor w/2 \rfloor, y - \lfloor w/2 \rfloor), \dots, p(x + \lfloor w/2 \rfloor, y + \lfloor w/2 \rfloor)\} \quad (7)$$

where $\lfloor \bullet \rfloor$ denotes the floor operation. First, the input image is split into a series of overlapping windows that have different sizes of patches, like a Gaussian pyramid, which is defined as follows:

$$w_t = \begin{cases} w_b & t = T \\ 2^{T-t} * w_b & otherwise \end{cases} \quad (8)$$

where T represents the total number of CNN per channel, $T = 3$ in this paper, and w_t ($t = 1, \dots, T$) is the base patch size of $CNN_1, CNN_2, \dots, CNN_T$. As shown in Fig. 4, we adjust the size of the large patches to the base patch size ($w_b \times w_b$). Fig. 4 shows the process of multi-scale patches extracted from the input image. Since all patches of extracting multi-scale aspects are adjusted to the same size, we use the same CNN structure as shown in Fig. 4.

In the process of testing, the patches are extracted in a similar method as with training and input into the CNN to obtain a feature map M_c , as shown in Fig.2. All of the feature maps M_c are then inter-fused to obtain a single feature

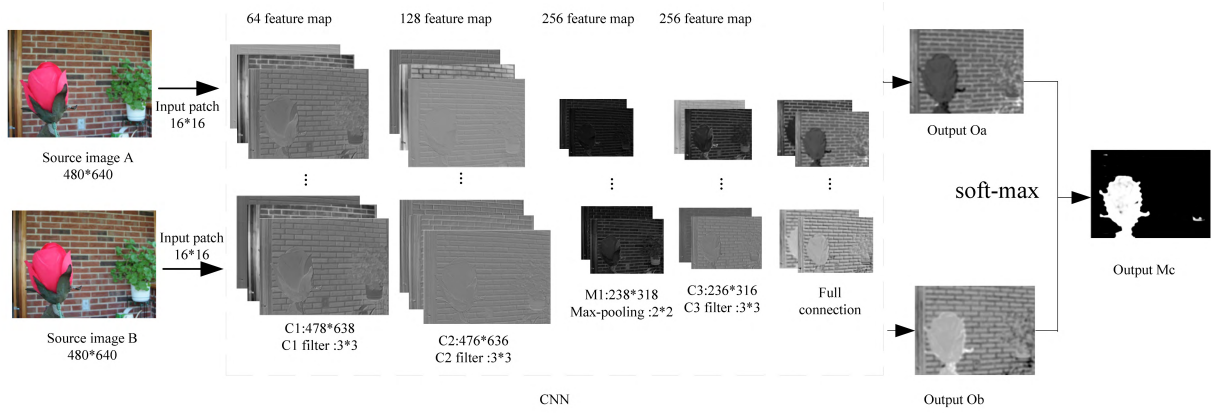


FIGURE 5. The CNN model used in the proposed method. The source image A and B are first subjected to multi-scale extraction; then, they are fed into the CNN method as shown in Fig. 3.

map M_f as shown in Fig. 2, which is defined as:

$$M_f = \sum_{c=1}^T \alpha_c M_c \quad (9)$$

where T indicates the total number of CNNs, and $\alpha_c = 1/T$ is a fusion parameter. Furthermore, the fused feature map M_f is post-processed using initial segmentation, morphological operations and the watershed transform. Post-processing aims to produce a segmentation map or decision map. First, the morphological operation is applied on M_f to remove the little black spots. Then, we use the watershed transform to generate the final segmentation, S_f , where the watershed lines would denote the region boundaries.

D. METHOD IMPLEMENTATION

The proposed CNN architecture in this paper is displayed in Fig. 5 and has three convolutional layers and one subsampling/max-pooling layer in the network.

- (1) A patch of 16×16 pixels is fed into the CNN.
- (2) The first and second convolution layer in the CNN can obtain 64 feature maps and 128 feature maps, respectively, by a 3×3 filter, and a stride of two convolutional layers is set to 1.
- (3) The filter size is set to 2×2 , and the stride of the max-pooling layer is set to 2, to obtain 256 feature maps.
- (4) The 256 feature maps are fed into another convolution layer to obtain 256 feature maps by a 3×3 filter.
- (5) The 256 feature maps are forwarded to be fully connected.
- (6) The output of the CNN is a 2-dimensional vector, and the 2-dimensional vector is composed of O_a and O_b in Fig. 5.

Further, a 2-way soft-max layer takes the 2-dimensional vector as input and obtains a probability distribution M_c over two classes [32]–[34], where, $M_c = e^{O_a} / (e^{O_a} + e^{O_b})$, that is, the processing of the soft-max layer. The base patch size w_b is the input of the CNN and is set to 16. Because all multi-scale patches are adjusted to the same base size, all of the CNNs in Fig. 2 share the same structure.

The overlapping patches from the input image are extracted with three different sizes: 16×16 , 32×32 , and 64×64 . Then, the two larger patches of the three different sizes, 32×32 and 64×64 , are downsized to 16×16 through a bicubic transformation. Therefore, the patches fed into the CNN are of the same size but reveal different contexts. When training, the patches are rotated at $\pm 90^\circ/10^\circ$ across the vertical and horizontal axes. The purpose of this step is to introduce invariance to such changes in the CNN. Through these pre-processing steps, the patches are fed into the CNN framework for training.

E. TRAINING

Similar to the CNN-based tasks [33], [34], the soft-max loss function is employed as the objective of the proposed CNN framework. In this article, stochastic gradient descent is employed to minimize the loss function. The weight decay and the momentum are set to 0.0005 and 0.9, respectively, in our CNN training procedure. The weights are updated using the following rule:

$$v_{i+1} = 0.9 * v_i - 0.0005 * \theta * w_i - \theta * \frac{\partial L}{\partial w_i} \quad (10)$$

where v is the momentum variable, θ is the learning rate, i is the iteration index, L is the loss function, and $\frac{\partial L}{\partial w_i}$ is the derivative of the loss function at w_i . In this paper, the proposed CNN framework use the prevalent deep learning framework Caffe [35]. The parameters of each convolutional layer in the CNN are initialized using the Xavier algorithm [36]. The biases in every convolutional layer are initially set to 0. The learning rate of all of the convolutional layers is equal and is initialized to 0.0001. When the loss reaches a steady state, we manually drop 10 times. Throughout the training process, the learning rate dropped once.

To better understand the MSCNN model, we offer a typical output feature map for each convolution layer. The source images A and B shown in Fig. 6 are employed as the inputs. The four corresponding feature maps of each convolution layer are shown in Fig. 6. From the first convolutional layers,

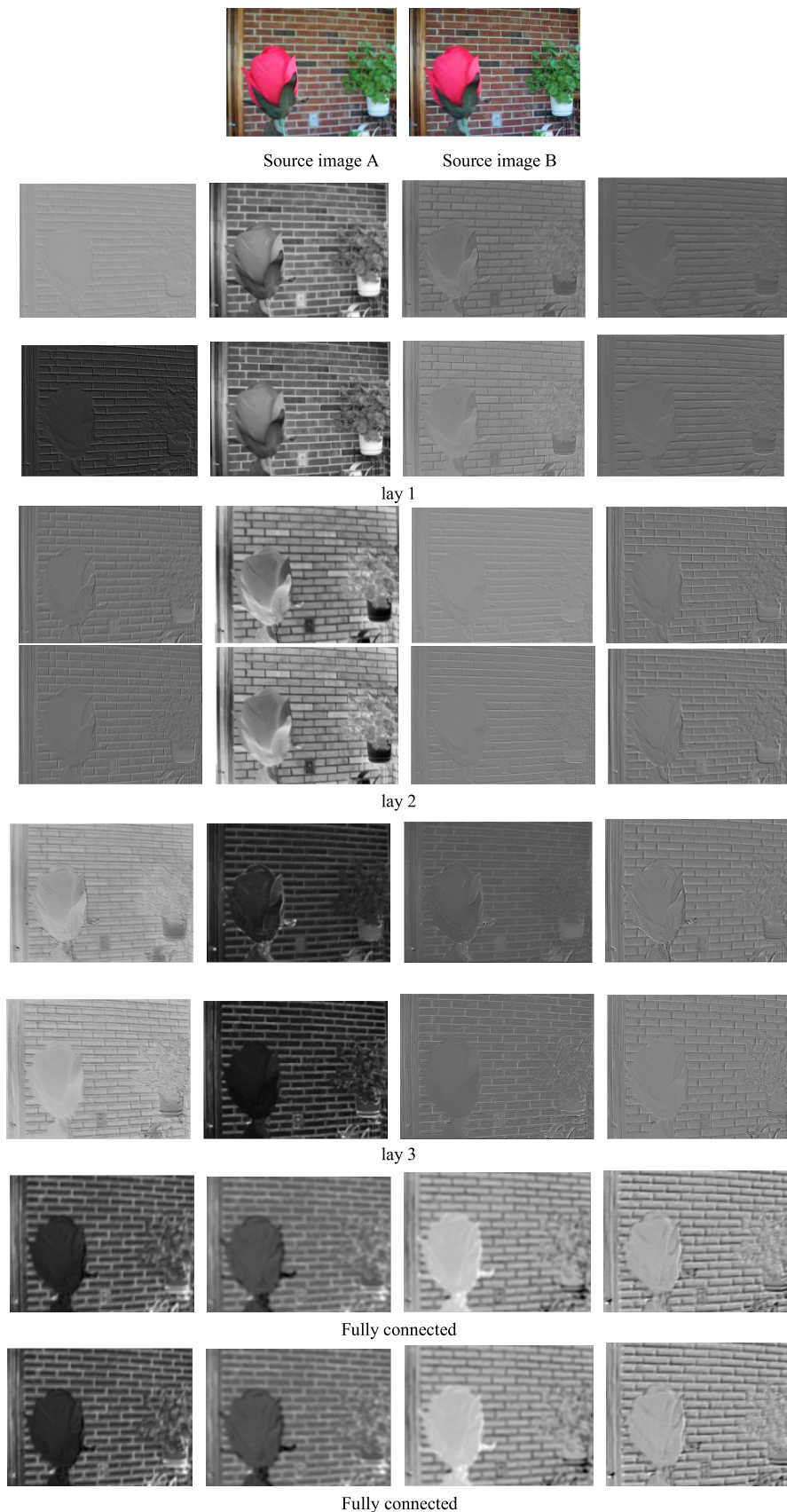


FIGURE 6. Some typical output feature maps of each convolutional layer. Here, "lay 1", "lay 2" and "lay 3" and the fully connected layer denotes the C1, C2 and C3 convolutional layers and the fully connected layer, respectively.

we can see that some feature maps catch the high frequency information of the input image as illustrated in the first column, third column and fourth column, while the second column is similar to the input images. This finding indicates that the first layer cannot adequately characterize the spatial details of the image. The feature maps obtained from the second layers are mainly focused on the extracted spatial details, covering different gradient directions. The output feature maps from the third convolutional layer successfully capture the focus information of the different source images. We can see that the feature map obtained by the fully connected layers shows that the focus area has been relatively clear.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we introduce the evaluation index system and analyse the experimental results as follows.

To verify the validity of the proposed MSCNN-based fusion algorithm, eight pairs of multi-focus images (including colour images and greyscale images) are used in our experiments. The proposed fusion method is compared with four state-of-the-art multi-focus image fusion methods, which are the MWGF [26], SSDI [44], CNN [37], and DSIFT [25]. The detailed analysis and discussions are given below.

A. COMPARISON WITH SEVERAL OTHER ALGORITHMS

We compare the validity of different fusion algorithms in terms of the subjectivity first. For this purpose, we mainly provide the “Lab” source image pair as an example to show the difference between the different methods. Fig. 7(c) shows the fused images obtained with several different fusion algorithms. In each of the fused images, the area around the boundary between the focused and defocused regions is magnified and displayed in the lower left corner. The CNN, MWGF and DSIFT based algorithms produce some undesirable artefacts in the fused image (as show in the right border of the clock), and the artefact is particularly pronounced for the CNN-based and MWGF-based methods. The fusion results based on the MWGF and SSDI fusion methods are blurred in the upper right corner of the clock.

To make a better comparison, Fig. 7(a) and (b) show the difference image obtained by subtracting source image A and source image B from each of the fused images, respectively, and the values of each difference image in Fig. 7 are normalized to the range of 0 to 1. The difference images CNN (b) and DSIFT (b) displayed in Fig. 7 clearly show that the CNN and DSIFT-based method has a partial residual in the upper right corner. The SSID-based approach is not sufficient in the integration of the head of the character. The difference image displayed in Fig. 7 SSDI (b) also reveals this limitation. According to Fig. 7 MWDF (b), one can observe that the MWDF-based approach performs well in terms of extraction details, except for the border area. In summary, the fusion image obtained by the proposed method has the highest visual quality in all five of these methods, which can be further proved by the difference image displayed in Fig. 7 (a) and (b).

The fused results of the ‘Temple’ image set are shown in Fig. 8. To clearly demonstrate the details of the fused results, partial regions of the fused results are magnified, as shown in Fig. 9. As seen, the fusion results of each algorithm can achieve the goal of the image fusion. However, different quality fused images were produced by different fusion algorithms, depending on the performances of the various fusion methods. MWGF-based methods produce a fusion image with blurring effects, such as the boundary between the focused and defocused parts around the stone lion (see Fig. 9 (c)). Compared with several other algorithms, we can clearly see that the erosion in stone lion is more serious for the boundary area obtained by the MWGF-based method. Thus, the MWGF-based algorithm often cannot achieve the ideal fusion image from the source image.

It can be seen from Fig. 9 (d) that there are many jagged phenomena in the lower right corner of the boundary area. At the same time, the fusion image on the left side of the stone shows two black spots(see Fig. 9 (d)), which indicates that the integration of the SSDI-based method is not sufficient. Similarly, we can see from Fig. 5 (e) that the fused image obtained by the DSIFT-based method has many jagged phenomena in the lower right corner of the boundary region. Fig. 9 (a) and (b) show that the fusion image of MSCNN and CNN-based appear very good, and the boundaries shown in Fig. 9 (a) and (b) are relatively smooth with respect to several other methods. However, compared with Fig. 9 (b), the contour of the boundary of Fig. 9 (a) is closer to the stone lion.

Finally, because of the superiority of the proposed methods, MSCNN could accurately find the multi-focus boundary between the focused and defocused parts and then obtain a more accurate decision map from the source images than from the other four fusion algorithms in this paper. The fusion image of MSCNN based model shows the satisfactory visual quality compared to several other algorithms.

B. DECISION MAP OF THE PROPOSED METHOD

Since the fused images are difficult to categorize between good and bad, to further prove the validity of the MSCNN model for multi-focus image fusion, we mainly compare the decision map that is produced by a variety of methods. According to the decision map, one can clearly see the advantages and disadvantages of various fusion methods. The comparison results of eight pairs of input source images are shown in Fig. 10. The “choose-max” strategy is employed as the binary segmentation approach of the proposed fusion algorithm to obtain a binary segmented map from the feature map (as shown in Fig. 3) with a fixed threshold. Thus, for multi-focus image fusion, the binary segmented map/decision map can be considered to be the actual output of our MSCNN model. From the binary segmented map of Fig. 3, we can conclude that the gained segmented maps of the MSCNN model are highly efficacious in that most pixels are classified correctly, which illustrates the effectiveness of the learned MSCNN model.

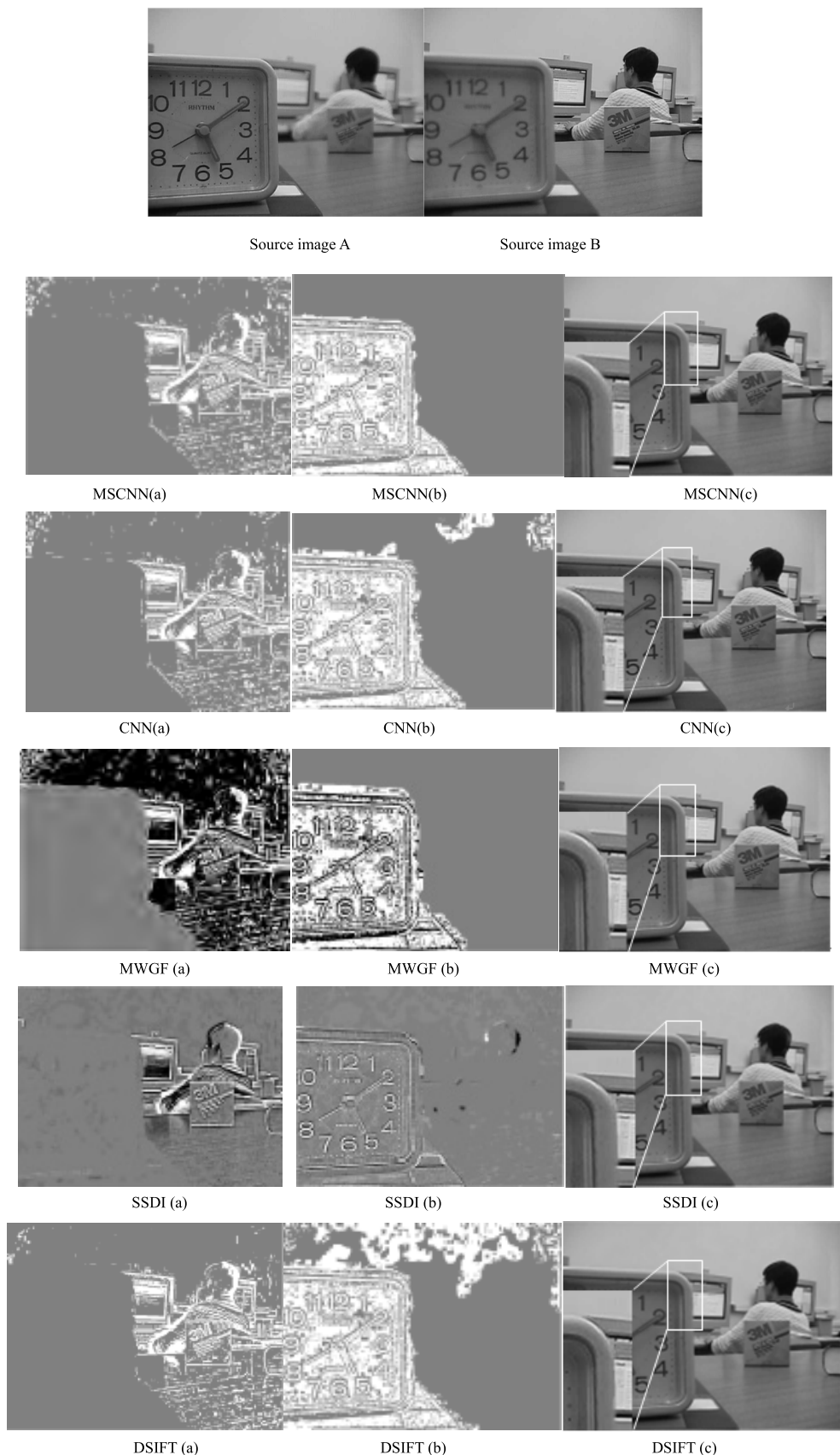


FIGURE 7. (a) and (b) show the difference image obtained by subtracting source image A and source image B from each fused image, respectively, and (c) shows the fused image.



FIGURE 8. The fusion results of all of the algorithms on the 'Temple' image set. (a) and (b) are the source image, (c)-(g) are the fusion results of MWGF, SSDI, CNN, DSIFT, and MSCNN; (h) clearly shows the boundary between the focused and defocused parts overlaid on the fusion image of MSCNN.



FIGURE 9. Magnified regions contain the boundary of the fusion results of all of the algorithms on the 'Temple' image set. (a)-(e) are the magnified regions extracted from the fused image through the MSCNN, CNN, MWGF, SSDI, and DSIFT-based methods.

However, there still exist some defects in light in the binary segmented map. First, a number of pixels are sometimes misclassified, which leads to the emergence of holes or small

regions in the segmented maps. Therefore, we use mathematical morphology to address the binary segmentation map and to obtain the final decision map. The final decision

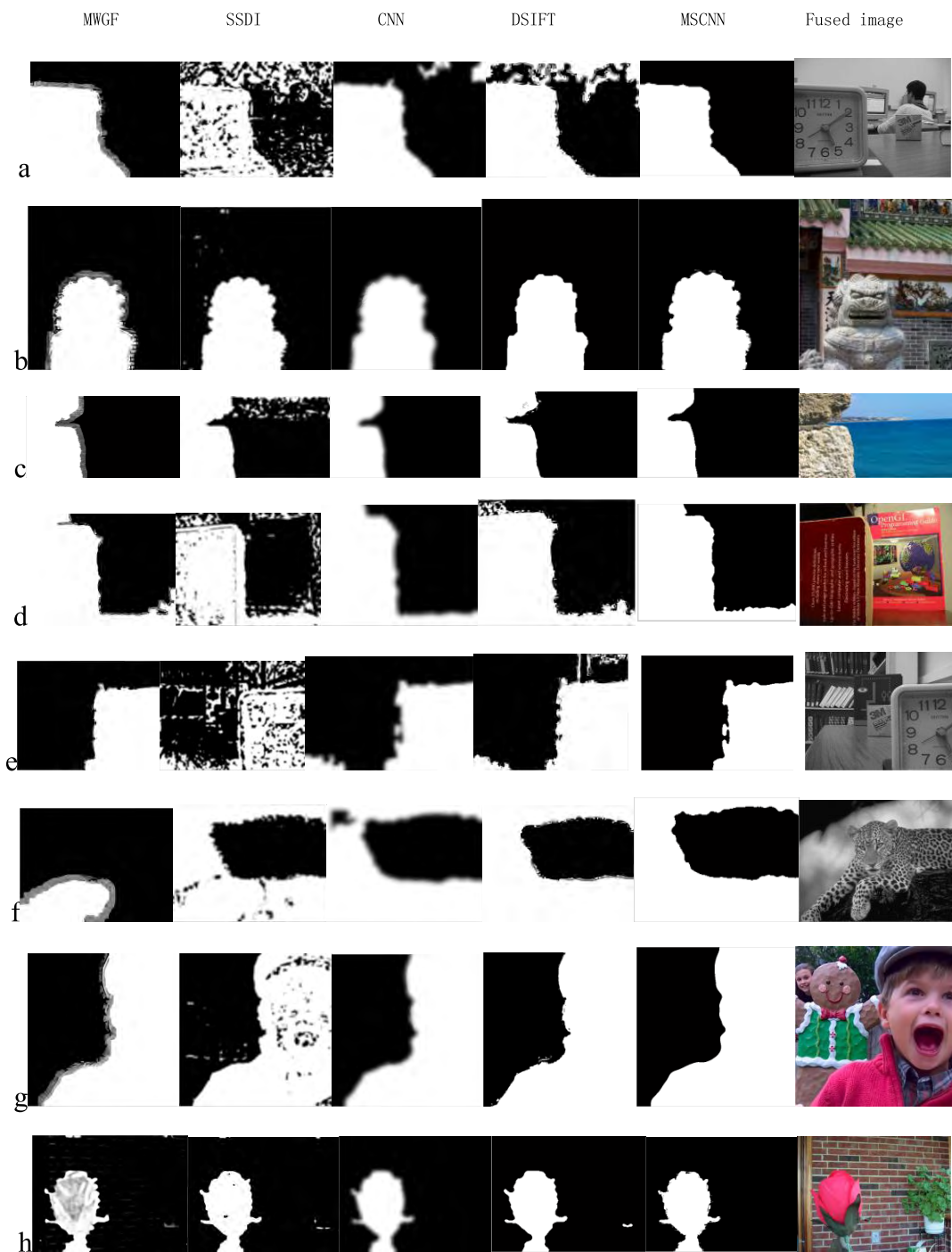


FIGURE 10. The decision map obtained by MWGF, SSDI, CNN, DSIFT and MSCNN. (a) Lab. (b) Temple. (c) Seascape. (d) Book. (e) Desk. (f) Leopard. (g) Children. (h) Flower.

maps displayed in the fifth column of Fig. 10 obtained from the MSCNN-based methods are very precise in the boundary (which has been proven to be correct in Fig. 9), which results in higher visual quality fusion results shown in the last column of Fig. 10.

C. OBJECTIVE CRITERIA OF THE PROPOSED METHOD

To prove the validity and practicability of the proposed algorithm, the three indexes of mutual information MI,

$Q^{AB/F}$ and $Q(A,B,F)$ are used as the objective evaluation index of information fusion performance [40]–[43]. $Q(A,B,F)$ is a similarity based quality metric [45] that is the objective evaluation index of information fusion performance based on the structural similarity (SSIM) metric [46] without requiring the reference image. The objective performance on the fused images using the five fusion methods are listed in table 1, from which we observe that the MSCNN-based method provides the best fusion results

TABLE 1. Comparison of objective criteria of different methods and multi-focus images.

		MWGF	SSDI	CNN	DSIFT	MSCNN
Lab	MI	8.0618	8.1412	8.6008	8.5201	8.8044
	$Q^{AB/F}$	0.7147	0.7528	0.7573	0.7585	0.7588
	Q(A,B,F)	0.8746	0.8823	0.8974	0.9132	0.9148
Temple	MI	5.9655	7.0896	6.8895	7.3514	7.4177
	$Q^{AB/F}$	0.7501	0.7634	0.7590	0.7643	0.7623
	Q(A,B,F)	0.8992	0.9125	0.9063	0.9138	0.9251
Seascape	MI	7.1404	7.4824	7.6285	7.9487	8.0214
	$Q^{AB/F}$	0.7059	0.7110	0.7113	0.7126	0.7122
	Q(A,B,F)	0.9366	0.9473	0.9481	0.9452	0.9547
Book	MI	8.2368	8.4008	8.7796	8.6623	8.8947
	$Q^{AB/F}$	0.7240	0.7260	0.7277	0.7134	0.7284
	Q(A,B,F)	0.9120	0.9221	0.9374	0.9045	0.9473
Leopard	MI	9.9474	10.8887	10.8792	10.9226	10.9420
	$Q^{AB/F}$	0.8175	0.8171	0.7973	0.8069	0.8267
	Q(A,B,F)	0.9435	0.9325	0.9218	0.9572	0.9748
Children	MI	8.2622	7.8505	8.3338	8.5252	8.5363
	$Q^{AB/F}$	0.6741	0.6799	0.7408	0.7394	0.7384
	Q(A,B,F)	0.8675	0.8752	0.9263	0.9255	0.9341
Flower	MI	8.3255	8.1049	8.2659	8.5365	8.6125
	$Q^{AB/F}$	0.6913	0.6490	0.7183	0.7159	0.7157
	Q(A,B,F)	0.9460	0.9207	0.9566	0.9479	0.9689

by considering the metrics MI and $Q(A,B,F)$. According to the scores of the metric $Q^{AB/F}$, one can observe that the MSCNN-based method provides the satisfactory fusion results for the “Lab”, “Book” and “Leopard” images, while the DSIFT outperforms the MSCNN-based method for the “Temple” and “Seascape” images, and the CNN outperforms the MSCNN-based method for the “Children” and “Flower” images. The above results indicate that MSCNN-based method needs to be improved in protecting the edge information in the fusion process, since the metric $Q^{AB/F}$ considers the fused image containing all the input edge information as the ideal fusion result.

V. CONCLUSIONS

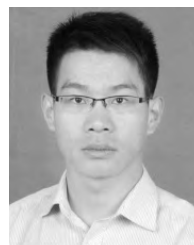
In this paper, we proposed a novel image segmentation-based multi-focus image fusion through a multi-scale convolutional neural network, in which the task of detecting the decision map is treated as image segmentation between the focused and defocused regions from the source images. The proposed method achieves segmentation through a multi-scale convolutional neural network (MSCNN), which conducts multi-scale analysis on each input image to derive the individual feature map on the region boundaries between the focused and defocused regions. The feature maps are then inter-fused to produce a fused feature map. Furthermore, the fused map

is post-processed using initial segmentation; morphological operation and the watershed transform to obtain the segmented map/decision map. We illustrate that the decision map obtained from MSCNN is trustworthy and that it can lead to high-quality fusion results. Experimental results evidently validate that the proposed algorithm can obtain optimum fusion performance in light of both qualitative and quantitative evaluations. But the max-pooling of feature mapping layer of traditional CNN, which is present in all modern CNN model used for dimensionality reduction of feature mapping, leads to the loss of information of the feature map due to the use of downsampling. How to efficiently avoid the loss of information may be a further development of the MSCNN-based method.

REFERENCES

- [1] T. Stathaki, *Image Fusion: Algorithms and Applications*. San Diego, CA, USA: Academic, 2008.
- [2] P. J. Burt and E. H. Adelson, “The Laplacian pyramid as a compact image code,” *IEEE Trans. Commun.*, vol. COM-31, no. 4, pp. 532–540, Apr. 1983.
- [3] A. Toet, “A morphological pyramidal image decomposition,” *Pattern Recognit. Lett.*, vol. 9, no. 4, pp. 255–261, May 1989.
- [4] H. Li, B. S. Manjunath, and S. K. Mitra, “Multisensor image fusion using the wavelet transform,” *Graph. Models Image Process.*, vol. 57, no. 3, pp. 235–245, May 1995.

- [5] J. J. Lewis, R. O'Callaghan, S. G. Nikolov, D. R. Bull, and N. Canagarajah, "Pixel- and region-based image fusion with complex wavelets," *Inf. Fusion*, vol. 8, no. 2, pp. 119–130, Apr. 2007.
- [6] Q. Zhang and B.-L. Guo, "Multifocus image fusion using the non-subsampled contourlet transform," *Signal Process.*, vol. 89, no. 7, pp. 1334–1346, 2009.
- [7] G. Piella, "A general framework for multiresolution image fusion: From pixels to regions," *Inf. Fusion*, vol. 4, no. 4, pp. 259–280, Dec. 2003.
- [8] X. Li, H. Li, Z. Yu, and Y. Kong, "Multifocus image fusion scheme based on the multiscale curvature in nonsubsampling contourlet transform domain," *Opt. Eng.*, vol. 54, no. 7, p. 073115, Jul. 2015.
- [9] Y. Liu, S. Liu, and Z. Wang, "A general framework for image fusion based on multi-scale transform and sparse representation," *Inf. Fusion*, vol. 24, no. 1, pp. 147–164, Jul. 2015.
- [10] N. Mitianoudis and T. Stathaki, "Pixel-based and region-based image fusion schemes using ICA bases," *Inf. Fusion*, vol. 8, no. 2, pp. 131–142, Apr. 2007.
- [11] B. Yang and S. Li, "Multifocus image fusion and restoration with sparse representation," *IEEE Trans. Instrum. Meas.*, vol. 59, no. 4, pp. 884–892, Apr. 2010.
- [12] J. Liang, Y. He, D. Liu, and X. Zeng, "Image fusion using higher order singular value decomposition," *IEEE Trans. Image Process.*, vol. 21, no. 5, pp. 2898–2909, May 2012.
- [13] Y. Jiang and M. Wang, "Image fusion with morphological component analysis," *Inf. Fusion*, vol. 18, no. 1, pp. 107–118, Jul. 2014.
- [14] Z. Liu, Y. Chai, H. Yin, J. Zhou, and Z. Zhu, "A novel multi-focus image fusion approach based on image decomposition," *Inf. Fusion*, vol. 35, pp. 102–116, May 2017.
- [15] W. Huang and Z. Jing, "Multi-focus image fusion using pulse coupled neural network," *Pattern Recognit. Lett.*, vol. 28, no. 9, pp. 1123–1132, Jul. 2007.
- [16] W. Huang and Z. Jing, "Evaluation of focus measures in multi-focus image fusion," *Pattern Recognit. Lett.*, vol. 28, no. 4, pp. 493–500, Mar. 2007.
- [17] S. Li, J. T. Kwok, and Y. Wang, "Combination of images with diverse focuses using the spatial frequency," *Inf. Fusion*, vol. 2, no. 3, pp. 169–176, Sep. 2001.
- [18] S. Li, J. T. Kwok, and Y. Wang, "Multifocus image fusion using artificial neural networks," *Pattern Recognit. Lett.*, vol. 23, no. 8, pp. 985–997, Jun. 2002.
- [19] V. Aslantas and R. Kurban, "Fusion of multi-focus images using differential evolution algorithm," *Expert Syst. Appl.*, vol. 37, pp. 8861–8870, Dec. 2010.
- [20] I. De and B. Chanda, "Multi-focus image fusion using a morphology-based focus measure in a quad-tree structure," *Inf. Fusion*, vol. 14, no. 2, pp. 136–146, Apr. 2013.
- [21] X. Bai, Y. Zhang, F. Zhou, and B. Xue, "Quadtree-based multi-focus image fusion using a weighted focus-measure," *Inf. Fusion*, vol. 22, no. 1, pp. 105–118, Mar. 2015.
- [22] M. Li, W. Cai, and Z. Tan, "A region-based multi-sensor image fusion scheme using pulse-coupled neural network," *Pattern Recognit. Lett.*, vol. 27, no. 16, pp. 1948–1956, 2006.
- [23] X. Bai, M. Liu, Z. Chen, P. Wang, and Y. Zhang, "Multi-focus image fusion through gradient-based decision map construction and mathematical morphology," *IEEE Access*, vol. 4, pp. 4749–4760, 2016.
- [24] S. Li, X. Kang, and J. Hu, "Image fusion with guided filtering," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2864–2875, Jul. 2013.
- [25] Y. Liu, S. Liu, and Z. Wang, "Multi-focus image fusion with dense SIFT," *Inf. Fusion*, vol. 23, pp. 139–155, May 2015.
- [26] Z. Zhou, S. Li, and B. Wang, "Multi-scale weighted gradient-based fusion for multi-focus images," *Inf. Fusion*, vol. 20, pp. 60–72, Nov. 2014.
- [27] Z.-H. Zhao, S.-P. Yang, and Z.-Q. Ma, "License plate character recognition based on convolutional neural network LeNet-5," *J. Syst. Simul.*, vol. 22, no. 3, pp. 638–641, 2010.
- [28] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.
- [29] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 10, pp. 1533–1545, Oct. 2014.
- [30] E. A. Smirnov, D. M. Timoshenko, and S. N. Andrianov, "Comparison of regularization methods for imagenet classification with deep convolutional neural networks," *AASRI Procedia*, vol. 6, no. 1, pp. 89–94, 2014.
- [31] Accessed on Jan. 1, 2017. [Online]. Available: https://en.wikipedia.org/wiki/Deep_learning
- [32] Y. Kim and T. Moon, "Human detection and activity classification based on micro-Doppler signatures using deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 1, pp. 8–12, Jan. 2016.
- [33] S. S. Farfade, M. J. Saberian, and L.-J. Li, "Multi-view face detection using deep convolutional neural networks," in *Proc. 5th ACM Int. Conf. Multimedia Retr.*, 2015, pp. 643–650.
- [34] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.
- [35] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 675–678.
- [36] S. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.
- [37] Y. Liu, X. Chen, H. Peng, and Z. Wang, "Multi-focus image fusion with a deep convolutional neural network," *Inf. Fusion*, vol. 36, pp. 191–207, Jul. 2017.
- [38] Y. Zhang, X. Bai, and T. Wang, "Boundary finding based multi-focus image fusion through multi-scale morphological focus-measure," *Inf. Fusion*, vol. 35, pp. 81–101, May 2017.
- [39] Z. Wang, S. Wang, and Y. Zhu, "Multi-focus image fusion based on the improved PCNN and guided filter," *Neural Process. Lett.*, vol. 45, no. 1, pp. 75–94, Feb. 2017.
- [40] Y. Yang, S. Tong, S. Huang, and P. Lin, "Multifocus image fusion based on NSCT and focused area detection," *IEEE Sensors J.*, vol. 15, no. 5, pp. 2824–2838, May 2015.
- [41] G. Piella and H. Heijmans, "A new quality metric for image fusion," in *Proc. IEEE Int. Conf. Image Process.*, Amsterdam, The Netherlands, Sep. 2003, pp. 173–176.
- [42] G. Bhatnagar, Q. M. J. Wu, and Z. Liu, "Directive contrast based multimodal medical image fusion in NSCT domain," *IEEE Trans. Multimedia*, vol. 15, no. 5, pp. 1014–1024, Aug. 2013.
- [43] B. Yang and S. Li, "Pixel-level image fusion with simultaneous orthogonal matching pursuit," *Inf. Fusion*, vol. 13, no. 1, pp. 10–19, Jan. 2012.
- [44] D. Guo, J. Yan, and X. Qu, "High quality multi-focus image fusion using self-similarity and depth information," *Opt. Commun.*, vol. 338, pp. 138–144, Mar. 2015.
- [45] C. Yang, J.-Q. Zhang, X.-R. Wang, and X. Liu, "A novel similarity based quality metric for image fusion," *Inf. Fusion*, vol. 9, no. 2, pp. 156–160, Apr. 2008.
- [46] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.



CHAOBEN DU received the B.S. and M.S. degrees from the Xinjiang University in 2009 and 2012, respectively. He is currently pursuing the Ph.D. degree with Northwestern Polytechnical University. His current research interests include image processing and fusion.



SHESHENG GAO is currently a Professor with the School of Automatics, Northwestern Polytechnical University, China. His research interests include control theory and engineering, navigation, guidance and control, and information fusion.

• • •