Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Multi-focus image fusion based on fully convolutional networks*

Rui GUO[1,2], Xuan-jing SHEN[1,2], Xiao-yu DONG[1,2], Xiao-li ZHANG[†‡1,2]

[1]*Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China*

[2]*College of Computer Science and Technology, Jilin University, Changchun 130012, China*

[†]E-mail: zhangxiaoli@jlu.edu.cn

Received July 7, 2019; Revision accepted Oct. 9, 2019; Crosschecked June 10, 2020

Abstract: We propose a multi-focus image fusion method, in which a fully convolutional network for focus detection (FD-FCN) is constructed. To obtain more precise focus detection maps, we propose to add skip layers in the network to make both detailed and abstract visual information available when using FD-FCN to generate maps. A new training dataset for the proposed network is constructed based on dataset CIFAR-10. The image fusion algorithm using FD-FCN contains three steps: focus maps are obtained using FD-FCN, decision map generation occurs by applying a morphological process on the focus maps, and image fusion occurs using a decision map. We carry out several sets of experiments, and both subjective and objective assessments demonstrate the superiority of the proposed fusion method to state-of-the-art algorithms.

Key words: Multi-focus image fusion; Fully convolutional networks; Skip layer; Performance evaluation
https://doi.org/10.1631/FITEE.1900336 **CLC number:** TP37

1 Introduction

Image fusion refers to the technology combining images of the same target collected by multiple sensors to obtain a higher-quality image (Saha et al., 2013; Chen and Qin, 2015). The existing algorithms can be divided into three levels from low to high: pixel, feature, and decision. Pixel-level fusion algorithms operate directly on the data obtained by sensors, and are closely related to super-resolution reconstruction (Kim and Kwon, 2010), image denoising (Huhle et al., 2010), and navigation (Bavirisetti and Dhuli, 2016). Pixel-level fusion is also the basis of feature- and decision-level fusion. Multi-focus image fusion is a hot topic in pixel-level fusion, and it has been widely used in cameras and microscopes (Li et al., 2017). It is well known that it is almost impossible for an optical imaging system to obtain an image in which all the objects at different distances are in focus. Multi-focus image fusion is a process that uses images with different focuses from the same scene to obtain an all-in-focus image (Zhang Q and Levine, 2016). At present, multi-focus image fusion can be roughly divided into two classes: transform-domain-based methods and spatial-domain-based methods (Li et al., 2013b).

Transform-domain-based algorithms are the ones in which source images are transformed into other spaces before information fusion. In recent years, many fusion methods in this class have been developed, e.g., the fusion method based on Laplacian pyramid transformation (Juočas et al., 2019), the wavelet analysis based method, curvelet transform, and (subsampled or non-subsampled) contourlet transform (Zhang BH et al., 2013; Yang et al., 2015). These recent algorithms have overcome some shortcomings in early transform-domain-based algorithms,

‡ Corresponding author
* Project supported by the National Natural Science Foundation of China (No. 61801190), the Natural Science Foundation of Jilin Province, China (No. 20180101055JC), and the Outstanding Young Talent Foundation of Jilin Province, China (No. 20180520029JH)
ORCID: Rui GUO, https://orcid.org/0000-0001-5246-0189; Xiao-li ZHANG, https://orcid.org/0000-0001-8412-4956
© Zhejiang University and Springer-Verlag GmbH Germany, part of Springer Nature 2020

such as pseudo-Gibbs phenomena and computational complexity problems. Zhou et al. (2014) proposed a multi-scale focus measurement method based on the image structure to determine the gradient weights for a multi-focus image fusion method. The fusion method based on morphological filtering proposed by Li et al. (2013a) can well solve the fusion problem of dynamic focus problems. Although the method based on the transform domain can effectively avoid the block effect, it cannot extract the pixels from the multi-focus images directly, causing the original image information to be less preserved. In addition, the fused image is often unacceptable because of artificial effect.

Compared with transform-domain-based algorithms, spatial-domain-based algorithms can retain more visual information from the original images. In addition, they are more efficient and easier to implement. These methods can be roughly divided into three categories: pixel-, block-, and region-based methods. For example, Saha et al. (2013) emphasized the relatively unique features of a source image with respect to the other source images so as to fuse multi-focus images. Huang and Jing (2007) proposed a block-based method by a pulse-coupled neural network (PCNN). Aslantas and Kurban (2010) proposed a novel optimal method for multi-focus image fusion using a differential evolution algorithm. Bai et al. (2015) proposed a new quadtree-based algorithm for multi-focus image fusion, in which the source images are decomposed into blocks with optimal sizes in a quadtree structure. However, spatial-domain-based methods always cause spectral distortion and block effect, so many other works are devoted to solving these problems.

Although much attention has been paid to multi-focus image fusion based on traditional algorithms, these algorithms have some defects and need further improvement. In recent years, there has been an increasing interest in methods based on convolutional neural networks (CNNs), which have been widely used in computer vision, and achieved excellent results in facial recognition, pose estimation, image classification, patch similarity analysis, semantic segmentation, etc. (Ayyalasomayajula et al., 2019; Ren et al., 2019; Wang et al., 2019; Zhu et al., 2019). Many results are much better than those obtained by the traditional methods, because the CNN for image feature extraction has a wider range of applications. Li et al. (2002) first used neural networks to handle multi-focus image fusion. The neural network in their algorithm is only equivalent to a classifier that formulates fusion rules. Recent experiments (Liu et al., 2017) regard one pair of image blocks with a size of 16×16 in a multi-focus image as one instance and then convert the multi-focus image fusion problem into a binary classification problem, using a CNN to judge the relationship of this pair of image blocks. Amin-Naji et al. (2019) deemed that the results of an ensemble of CNNs are better than those of one single CNN. Hence, they proposed an ensemble learning based CNN for multi-focus image fusion. Mustafa et al. (2019) proposed a complete unsupervised end-to-end trainable deep CNN model for multi-focus image fusion. Zhao et al. (2019) proposed a novel end-to-end deep learning based approach for multi-focus image fusion with a natural enhancement, which combines multi-level visually distinctive features.

In this paper, we propose a fully convolutional network (FCN) for focus detection (FD-FCN) in the spatial space. Based on the detection results, pixels in focused regions are selected to construct an all-in-focus image. To improve the accuracy of the results, we combine multi-scale information in the network structure. Compared with traditional image fusion methods, the proposed method based on a CNN is simpler and easier to understand. It does not require artificially designing complex fusion rules or selecting thresholds. Experimental results demonstrate that our method has strong robustness, and can effectively avoid block effects, artificial effects, and other issues. In addition, the FCN can process source images with an arbitrary size. The main contributions of this study can be summarized as follows:

1. A new FCN with skip layers, denoted as FD-FCN, is proposed for focus detection.

2. A new training dataset for FD-FCN is constructed based on dataset CIFAR-10.

3. The algorithm can be easily extended to a fusing image sequence.

2 Fully convolutional network

In the field of computer vision, CNNs are extensively adopted for image classification, object detection, and so on. The CNN was proposed by

LeCun and Bengio (1995), and it can be seen as an extended version of traditional artificial neural networks. However, CNN's structure is more complex, because its capabilities are stronger, and the number of parameters is greatly reduced. In general, a CNN is a feedforward neural network with multiple hidden layers, making it capable of extracting global features from images.

The basic components of a CNN are convolutional layers, pooling layers, fully connected layers, and output layers. Taking the most general classification network as an example, after several convolutional operations to extract features, its output is a vector through the fully connected layer, and the output of the fully connected layer is fed to the softmax layer. Finally, we obtain a vector whose values correspond to the probability of belonging to each category, and the maximum value of this category is the final result.

However, in a trained CNN, except for the number of feature maps in each layer, the number of neurons in the fully connected layer cannot be changed. This leads to the fact that the size of the feature maps cannot be changed when the feature maps feed into the fully connected layer, although the convolutional operation does not limit the size of input images. In other words, in the process of training and testing, the size of the input image must be the same, which makes it difficult to apply the CNN to many fields, such as semantic segmentation and image fusion.

To solve this problem, Shelhamer et al. (2017) invented the FCN based on the CNN. The main contribution of the new network is to remove the fully connected layers. Without the restriction of fully connected layers, it will obtain a matrix with the same number of layers as classes, and this network can handle arbitrarily sized images. This feature makes FCNs be widely used for semantic segmentation problems.

The aim of pixel image fusion is to generate a more informative image with source images of the same size, and the size of source images is not fixed in practical applications. FCN makes it possible to deal with image fusion using deep learning. Hence, we propose a multi-focus image fusion algorithm using an FCN. In the algorithm, the output of the network is the focus detection result. As the model learns on large-scale datasets, it can generate more precise focus detection results than traditional methods in theory.

3 The proposed fusion algorithm

3.1 Framework

In a multi-focus image fusion algorithm, the key issue is to detect focused regions. However, it is still an open problem. In this study, we detect focused regions using the theory of deep learning. We assume that two source images, denoted as $\boldsymbol{A}$ and $\boldsymbol{B}$, are well registered. A new FCN is designed to obtain a focus map. The overall flow diagram of the proposed method (Fig. 1) can be summarized as follows:

1. Input the source images $\boldsymbol{A}$ and $\boldsymbol{B}$ into our trained FD-FCN to obtain the focus maps (score maps) of $\boldsymbol{A}$ and $\boldsymbol{B}$, which are denoted as **SA** and **SB**, respectively.

2. Conduct a series of morphological processing steps on focus maps to obtain the final decision map **Ds**.

3. Based on **Ds**, the fused image can be obtained by $\boldsymbol{F}=\mathbf{Ds}\times\boldsymbol{A}+(\mathbf{1}-\mathbf{Ds})\times\boldsymbol{B}$.

The network structure of the proposed FD-FCN is presented in Section 3.2. The training process is shown in Section 3.3. The morphological processing is discussed in Section 3.4.

3.2 Network structure of FD-FCN

As shown in Fig. 1, the FD-FCN we established is an FCN with skip layers, capable of merging multi-scale features. This network is composed mainly of convolutional layers, pooling layers, and upsampling layers. After a series of operations, the network outputs a map that is half the size of the source image. Then, we will create a focus map that has the same size as the source image using bilinear interpolation. In the focus map, each element represents the probability of the corresponding pixel belonging to focused regions. Three aspects of this structure should be specified:

1. FCN

It can be seen from Fig. 1 that the first layer of the network is the data input layer. The main structure consists of three modules, and each module includes mainly three convolutional layers and one pooling

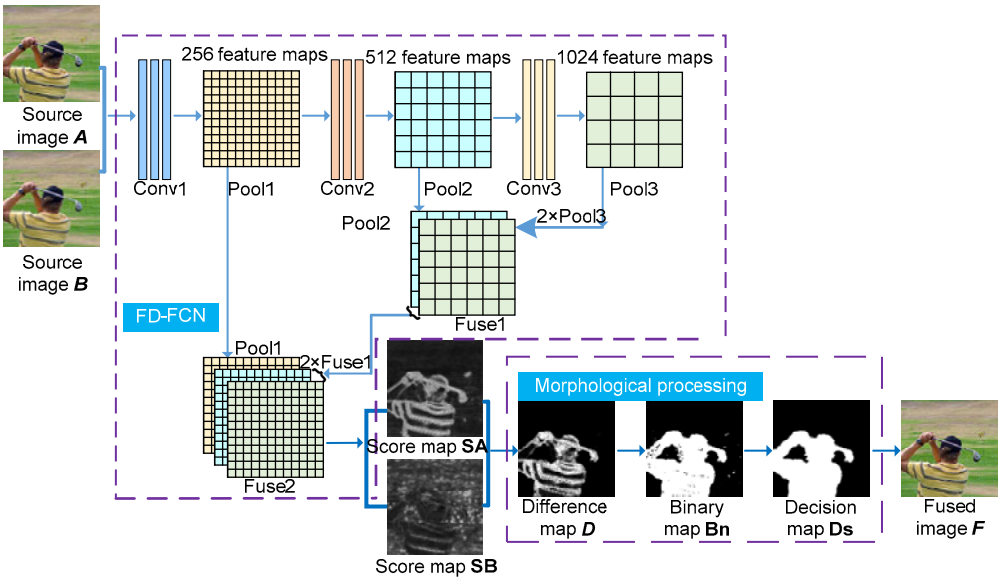

Fig. 1 Overall flow diagram of the proposed method

layer. The convolutional layer is the core of FD-FCN, used to extract features. As the size of the training dataset grows, each neuron will learn to filter out corresponding features. In FD-FCN, the size of the convolutional kernels in the convolutional layer is 3×3. This process of convolution can be expressed as

$$z = \boldsymbol{W} \otimes \boldsymbol{x} + \boldsymbol{b}, \tag{1}$$

where $\otimes$ is the convolutional operator, $\boldsymbol{W}$ and $\boldsymbol{b}$ represent the weight and bias matrices respectively, and $\boldsymbol{x}$ and z represent the input and output of the convolution respectively.

The pooling layer is used for downsampling feature maps. This process can further reduce the number of parameters and avoid overfitting. The pooling methods include mean pooling, maximum pooling, etc. Since maximum pooling has been widely used, all those used in FD-FCN are set as maximum pooling, that is, selecting the largest element in the area.

Taking the training process as an example, the size of images in the training set used in our experiments is 32×32. After the first module, the size of the feature map becomes 16×16, and then reduces to 8×8 and 4×4. In the end, we obtain an original map that is half the size of the original image (16×16), and will obtain a map that has the same size as the source image by bilinear interpolation. The map expresses the probability of the pixel in the source image belonging to focused regions.

Note that the model uses a smaller convolutional kernel, and it has been demonstrated that the 3×3 convolutional kernel on small images converges faster and that the performance is significantly improved compared to larger convolutional kernels. In addition, a 1×1 convolutional kernel is used for dimensionality reduction and to improve the network expressiveness. We use the ReLU activation function, which is sparser than the sigmoid function. ReLU not only avoids the vanishing gradient, but also allows the model to converge as quickly as possible. The function is defined as

$$\mathrm{ReLU}(x) = \begin{cases} x, & x \geq 0, \\ 0, & x < 0. \end{cases} \tag{2}$$

2. Multi-scale feature combinations

As shown in Fig. 1, we add skip layers in the structure. In FD-FCN, we will operate several downsamplings, and consequently the resolution of the image will decrease. For convolutional networks, shallow convolutional layers have a smaller perceptual domain and extract more local features, while deep convolutional layers with a larger perceptual domain can extract more abstract features. Due to the characteristics of multi-focus image fusion, we need both detailed and abstract features. It is difficult to

deal with the pixels around the boundaries of focused and defocused regions using just the abstract information preserved in deep layers. To process these regions more precisely, we need more detailed features. Therefore, we combine different layers' feature maps to obtain a multi-scale score map. The final score map combines different scales for more precise results. As shown in Fig. 1, Pool3 is upsampled and then merged with Pool2 to obtain Fuse1, and Fuse1 is upsampled and then merged with Pool1 to obtain Fuse2. After the last upsampling, the score map combined with multi-scale results has the same size as the source images.

3. Shallower network structure

In other image processing fields, source images are often downsampled to a very small size, and then upsampled to obtain the final result. Taking the FCN-32 in semantic segmentation as an example, the original images with size 32×32 are reduced to 1×1 after several downsamplings, and then are upsampled to generate the prediction map. When dealing with multi-focus image fusion problems, we find that the results of more downsampling operations are rough, but the feature maps obtained using fewer downsampling operations can better determine the clarity of each pixel. Therefore, the FD-FCN performs only three downsamplings.

3.3 Training dataset

To obtain better fusion results, we need a large and effective training set. However, there are no such datasets in the multi-focus image fusion field. It is a good choice to construct a dataset that is suitable for training the FD-FCN from an existing dataset in other fields. CIFAR-10, which has been widely used in object recognition, is selected as the base dataset to construct our object dataset. Compared with other datasets, the data features of CIFAR-10 are more complex. The dataset contains 10 categories, and each category has 6000 color images with size 32×32. In addition, these color images are independent of each other without overlapping.

When generating a training dataset, we take the original image as the focused region. At the same time, Gaussian blur is applied to all images with a variance of 15, simulating the defocused region. After these processes, our training dataset eventually includes 60 000 clear images and 60 000 blurred images. Because the images in the CIFAR-10 dataset are from the actual image, they contain a lot of information.

In some cases, to obtain more detailed information, we need to obtain information about the pixel-level images. There are many pixel-level label algorithms (Shelhamer et al., 2017) in some related research areas. We also use pixel-level labels in the experiment to obtain a more accurate clarity score for each pixel in the image. Although the size of the image used for training is definite, we can control the scale of the network according to the size of the labels, and find that when the size of labels is set to 16×16, the experiment performs well. So, in our experiment, the label of unprocessed clear images which are used to indicate the "focused region" is a 16×16 matrix full of 1, while the label of Gaussian-blurred images indicating the "defocused region" is a 16×16 matrix full of 0. In the training process, unlike the single-label problem, we calculate the loss value of the label matrix and network output by applying the mean square error (MSE), which has been widely used in regression problems.

When training the FD-FCN, each focused image and its blurred version are input into the network, and the network generates focus maps for the two images separately. Elements in each focus map are in [0, 1]. The focused regions tend to generate elements close to 1 in the focus map, and the defocused regions tend to generate elements close to 0 in the focus map. Hence, when testing, the framework can generate a focus map according to the real distribution of the focused and defocused regions.

3.4 Morphological processing

As shown in Fig. 2, in general, we select pixels by comparing scores in score maps. In the experiment, we make a difference map $\boldsymbol{D}$ between two score maps.

$$\boldsymbol{D} = \mathbf{SA} - \mathbf{SB}, \tag{3}$$

where **SA** and **SB** are the focus maps (score maps) of source images $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively.

$$\mathrm{Bn}(x, y) = \begin{cases} 1, & D(x, y) \geq 0, \\ 0, & D(x, y) < 0, \end{cases} \tag{4}$$

where **Bn** is the binary map.

Inevitably, there are some small closed region errors in our algorithm. We deal with these errors to improve the quality of the fused image. As shown in Fig. 2, we remove these small holes in the binary image **Bn**. Then we refine **Bn** using a guided filter (He et al., 2013), which can transfer the structures of **Bn** to **Ds**; that is to say, we use it to smooth the edges:

$$\mathbf{Ds} = \mathrm{GF}(\mathbf{Bn}, r, e), \tag{5}$$

where r is the local window radius and e the regularization parameter. As shown in Fig. 2, we finally obtain a decision map **Ds**, and then obtain the fused image $\boldsymbol{F}$ according to

$$F(x,y) = \mathrm{Ds}(x,y)A(x,y) + \left[1 - \mathrm{Ds}(x,y)B(x,y)\right]. \tag{6}$$

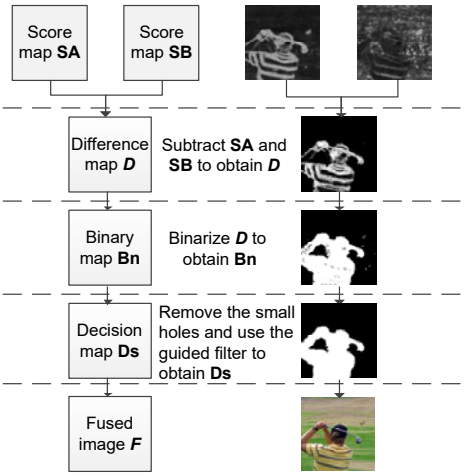

Fig. 2 Structure of morphological processing and a demonstration example

4 Experimental results and discussion

4.1 Focus detection results and fused images

We performed a series of experiments to examine the focus maps generated by FD-FCN and decision maps. Fig. 3 shows 10 sets of score maps, decision maps, and fused images generated by our algorithm, where (1)–(5) are color images, and (6)–(10) are gray images. Note that (3)–(5) and (8)–(10) in Fig. 3 are unregistered or dynamic images. For example, in (4), the sizes of the lions in the two source images are different, so are the backgrounds. The girl in (9) has different postures. Compared to the static multi-focus image that has been registered, research on the fusion of unregistered and dynamic images is lacking. Because our algorithm directly selects pixels from the source image based on the clarity of different pixels, it avoids the contamination of information in the blurred area, and will not produce artifacts or other bad effects.

4.2 Comparison experiments

4.2.1 Experimental settings

To examine the superiority of the proposed algorithm, we chose six state-of-the-art algorithms for comparison, including the image matting fusion (IMF) algorithm (Li et al., 2013b), discrete wavelet transform (DWT) based algorithm (Yang et al., 2014), multi-scale weighted gradient (MWG) based algorithm (Zhou et al., 2014), nonsubsampled contourlet transform (NSCT) based algorithm (Yang et al., 2015), guided filtering (GF) based algorithm (Li et al., 2013a), and CNN-based algorithm (Liu et al., 2017).

The existing methods for evaluating the performance of image fusion are generally classified into two categories: subjective and objective. Subjective methods refer to the way by organizing humans to evaluate the visual quality of fused images; this kind of methods is relatively reliable because the evaluation results are in accordance with humans' visual perception. However, a harsh experimental environment is required. Objective methods predict the visual quality of fused images by modeling a human visual system (HVS). They can fully avoid the drawbacks of subjective methods. In our experiments, the proposed fusion algorithm is compared with six state-of-the-art algorithms using both subjective and objective evaluation methods. The objective measures adopted in the experiments are as follows:

1. Mutual information (MI) (Qu et al., 2002)

The measure MI is defined as

$$\mathrm{MI}_{AF} = \sum_{i=0}^{L-1}\sum_{j=0}^{L-1} p_{AF}(i,j)\log_2 \frac{p_{AF}(i,j)}{p_A(i)p_F(j)}, \tag{7}$$

where p_A and p_F are the normalized histograms of source image $\boldsymbol{A}$ and fused image $\boldsymbol{F}$ respectively, and p_{AF} is the normalized joint histogram of $\boldsymbol{A}$ and $\boldsymbol{F}$. The MI_{BF} between fused image $\boldsymbol{F}$ and another source image $\boldsymbol{B}$ can be calculated in the same way. We set $\mathrm{MI}=\mathrm{MI}_{AF}+\mathrm{MI}_{BF}$. MI is used to calculate the amount of

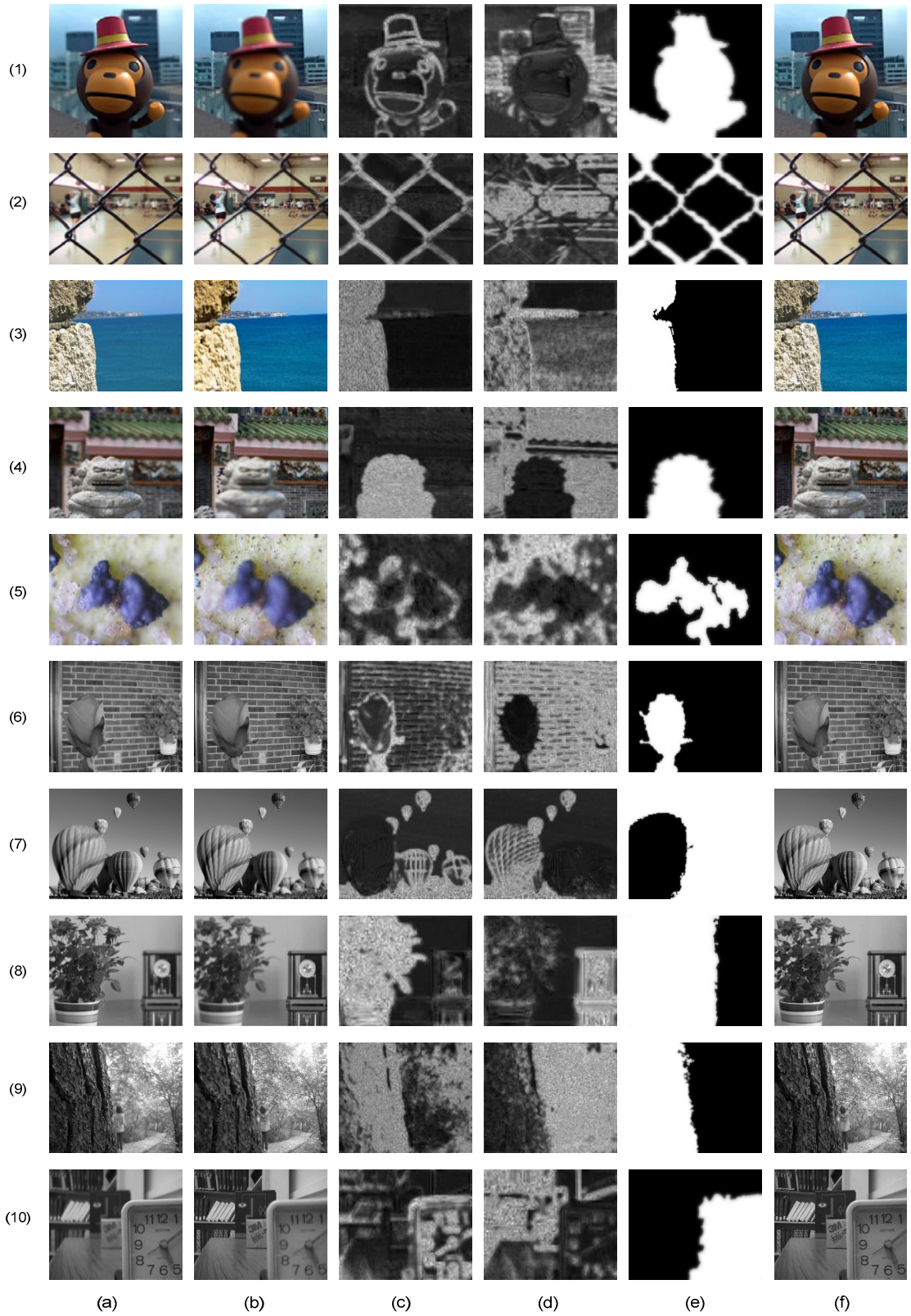

Fig. 3 Ten sets of examples: (a) source image *A*; (b) source image *B*; (c) score map SA; (d) score map SB; (e) decision map Ds; (f) fused image *F*

information that $\boldsymbol{F}$ retains from $\boldsymbol{A}$ and $\boldsymbol{B}$. The larger the MI, the better the fusion quality.

2. Edge information preservation (Q_{AB}^{F}) (Xydeas and Petrovic, 2000)

The measure Q_{AB}^{F} is defined as

$$Q_{AB}^{F}=\frac{\sum_{i=1}^{M}\sum_{j=1}^{N}\left[Q^{AF}(i,j)\omega^{A}(i,j)+Q^{BF}(i,j)\omega^{B}(i,j)\right]}{\sum_{i=1}^{M}\sum_{j=1}^{N}\left[\omega^{A}(i,j)+\omega^{B}(i,j)\right]}, \tag{8}$$

where Q^{AF} and Q^{BF} denote the similarities of the fused image $\boldsymbol{F}$ with the two source images $\boldsymbol{A}$ and $\boldsymbol{B}$ on the edge information respectively, and ω^{A} and ω^{B} are weights of Q^{AF} and Q^{BF} respectively. The measure Q_{AB}^{F} evaluates the quality of the fused image by calculating the amount of edge information transferred from the source images to the fused image, which can be considered an evaluation measure of spatial structure similarity.

3. Piella measure (Q_{e}) (Piella and Heijmans, 2003)

Based on the structural similarity model, Piella and Heijmans (2003) proposed an evaluation measure representing the significant information that is transferred from the source image to the fused image. The global evaluation index was obtained by weighting the summation of each small window, and Q_{e} is an index of edge fusion quality evaluation.

$$Q_{\mathrm{e}}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{F})=Q_{w}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{F})^{1-\alpha}\,Q_{w}(\boldsymbol{A}',\boldsymbol{B}',\boldsymbol{F}')^{\alpha}, \tag{9}$$

where $\boldsymbol{A}'$, $\boldsymbol{B}'$, and $\boldsymbol{F}'$ are the edge images of images $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{F}$ respectively, and α is a parameter balancing the two factors. We set α=0.5 in the experiments. $Q_w(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{F})$ calculates the local similarity of images $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{F}$, and the details can be found in Piella and Heijmans (2003).

4. Radon Wigner-Ville-based blur metric (Saleem et al., 2011)

Based on renyi entropy, Saleem et al. (2011) proposed a new image quality metric, defined as

$$Q_{\mathrm{R}}=\frac{1}{N}\sum_{i=1}^{N}R_{3}(\boldsymbol{F}_{i}), \tag{10}$$

where N is the total number of image Radon profiles, $\boldsymbol{F}_i$ is the ith Radon profile of fused image $\boldsymbol{F}$, and $R_3(\boldsymbol{F}_i)$ is expressed as

$$R_{3}(\boldsymbol{F}_{i})=-\frac{1}{2}\log_{2}\left(\sum_{n}\sum_{k}W_{F}^{3}(n,k)\right).$$

Here,

$$W_{F}(n,k)=2\underset{m\to k}{\mathrm{DFT}}\{\boldsymbol{F}[n+m]\times\boldsymbol{F}[n-m]\},\ m\in\langle N\rangle,$$

where DFT means discrete Fourier transform and <N> means any set of N consecutive integers. According to its theory, the value of Q_{R} decreases with increasing blur in fused images.

5. Visual information fidelity for fusion (VIFF) (Han et al., 2013)

VIFF measures the quality of fused images from the perspective of information sharing. VIFF establishes a link between visual indicators and image information, and is an indicator that has been presented in recent years with high complexity and good results. The measure contains four stages:

(1) Source and fused images are filtered and divided into blocks.

(2) Visual information is evaluated with or without distortion information in each block.

(3) The VIFF of each sub-band is calculated as follows:

$$\mathrm{VIFF}_{k}(\boldsymbol{I}_{F})=\frac{\sum_{b}\mathrm{FVID}_{k,b}(\boldsymbol{I}_{1},\boldsymbol{I}_{2},\ldots,\boldsymbol{I}_{n},\boldsymbol{I}_{F})}{\sum_{b}\mathrm{FVIND}_{k,b}(\boldsymbol{I}_{1},\boldsymbol{I}_{2},\ldots,\boldsymbol{I}_{n},\boldsymbol{I}_{F})}, \tag{11}$$

where $\mathrm{FVID}_{k,b}(\boldsymbol{I}_1, \boldsymbol{I}_2, \ldots, \boldsymbol{I}_n, \boldsymbol{I}_F)$ and $\mathrm{FVIND}_{k,b}(\boldsymbol{I}_1, \boldsymbol{I}_2, \ldots, \boldsymbol{I}_n, \boldsymbol{I}_F)$ are visual information of the fusion with and without distortion at the kth level, b^{th} block, respectively.

(4) The overall quality measure is determined by weighting the VIFF of each sub-band:

$$\mathrm{VIFF}(\boldsymbol{I}_{1},\boldsymbol{I}_{2},\ldots,\boldsymbol{I}_{n},\boldsymbol{I}_{F})=\sum_{k}p_{k}\mathrm{VIFF}_{k}(\boldsymbol{I}_{F}), \tag{12}$$

where p_k are the weighting coefficients (with k being a positive integer, k=4 in the experiment), set to be [0.465, 0, 0.070, 0.465] according to the experimental results in Han et al. (2013). Obviously, a larger VIFF indicates a better fusion performance.

In the proposed algorithm, there are two parameters: local window radius r and regularization parameter e. To determine the optimal values for these two parameters, we ran the proposed algorithm with different settings on the test set, and calculated the MSE between the generated decision map and ground truth. Fig. 4 shows the averaged MSE with different parameter settings. It reveals that the algorithm with r=8 and e=0.3^2 can generate the best decision map generally.

4.2.2 Comparative experiments

The proposed algorithm was compared with six state-of-the-art algorithms. Because of the space limitation, we list only four test image sets, which cover gray-gray and color-color image fusion.

Fig. 5 shows a set of fused color images named "Golf." Because this set of source images has more detailed information, the fusion algorithms' ability to deal with details can be well compared. Figs. 5a and 5b display the source images with different focuses. Figs. 5c–5i show the images fused by the different algorithms. The red and yellow boxes are at the borders between focused and defocused regions with more detailed information. These two areas can better evaluate the quality of the fused image. In general, although the IMF method retained the detailed information better, there was a lot of noise in the fused image. These patches also appeared in the shoulder area that we have marked with red rectangles; although the small ball in the corner of the red box was well preserved, the obvious spots of the whole image affected the overall effect. DWT also retained the details, but the image had block effects. MWG made a big mistake in the red box. NSCT, GF, and CNN worked well overall, but the details of the balls in the red and yellow boxes did not show up, such as the small ball in the yellow box next to the stick. Table 1

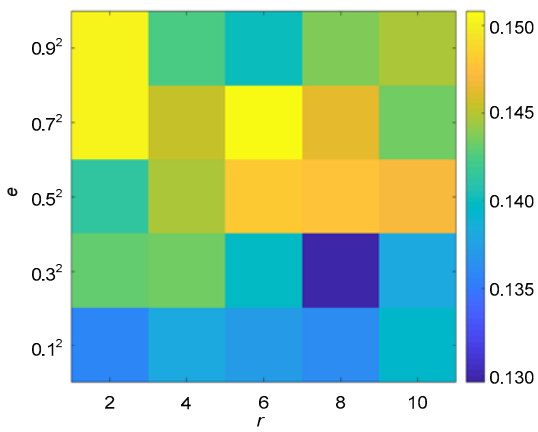

Fig. 4 Averaged mean square error (MSE) between the generated decision map and ground truth

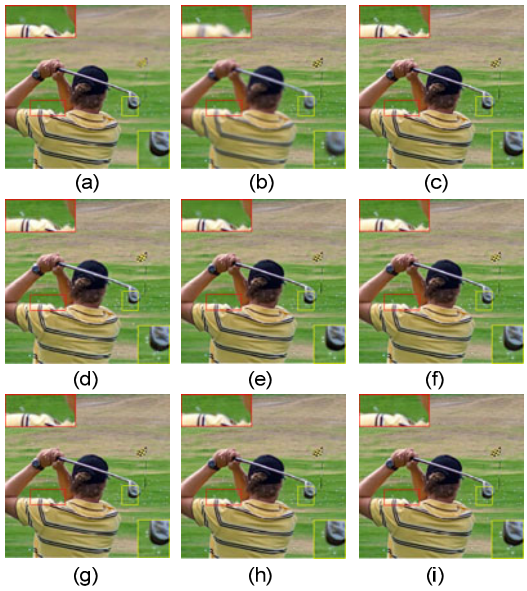

Fig. 5 The first set of colored images "Golf": (a) source image *A*; (b) source image *B*; (c) fused image by IMF; (d) fused image by DWT; (e) fused image by MWG; (f) fused image by NSCT; (g) fused image by GF; (h) fused image by CNN; (i) fused image by FD-FCN

References to color refer to the online version of this figure

Table 1 Quantitative evaluation of Fig. 5

Algorithm	MI	Q_{AB}^{F}	Q_e	Q_R	VIFF
IMF	7.199 840	0.741 351	0.789 912	1.742 883	0.584 032
DWT	5.312 529	0.683 448	0.792 999	1.555 027	0.490 717
MWG	7.080 758	0.740 648	0.800 766	1.691 695	0.611 451
NSCT	6.435 876	0.745 110	0.802 505	1.699 236	0.588 021
GF	7.180 822	0.754 383	0.801 949	1.734 907	0.610 761
CNN	7.395 021	0.754 614	0.801 807	1.745 298	0.613 490
FD-FCN	**7.556 523**	**0.755 076**	**0.802 554**	**1.753 202**	**0.616 000**

The best results are in bold

shows the quantitative evaluation in Fig. 5. The best results are marked in bold. From the table, it can be seen that the proposed algorithm gives the largest quality indexes for all the objective measures.

Fig. 6 shows a set of fused colored images named "Child." IMF also generated some unexpected spots in the fused images at the edge of the shoulder. DWT still had an obvious block effect. NSCT produced obvious artifacts around the boundaries of the objects. By magnifying the border between the focused and defocused areas, we can see that our fused image was clearer than others. Table 2 shows the quantitative evaluation of "Child," from which we can see that our method had better performance.

Fig. 7 shows the source and fused images named "Clock." To examine the details of the fused images, we amplified the region marked with a red box in each fused image. We can see from the figure that "3" was blurred and there was some noise in the fused image by IMF. Although the DWT algorithm preserved detailed visual information, there were obvious block effects. MWG and NSCT generated obvious distortions around the digit "8" in the red box. Compared to other images, some noise can be seen in GF. Both CNN and FD-FCN performed well, but the quantitative evaluation in Table 3 indicates that FD-FCN generated a fused image with higher visual quality.

Fig. 8 is also a set of classic multi-focus images named "Peisi." There were still spots in the IMF, and the edge of the table was blurred. DWT still had obvious block effects. MWG and GF did not perform

Fig. 6 The second set of color images "Child": (a) source image *A*; (b) source image *B*; (c) fused image by IMF; (d) fused image by DWT; (e) fused image by MWG; (f) fused image by NSCT; (g) fused image by GF; (h) fused image by CNN; (i) fused image by FD-FCN

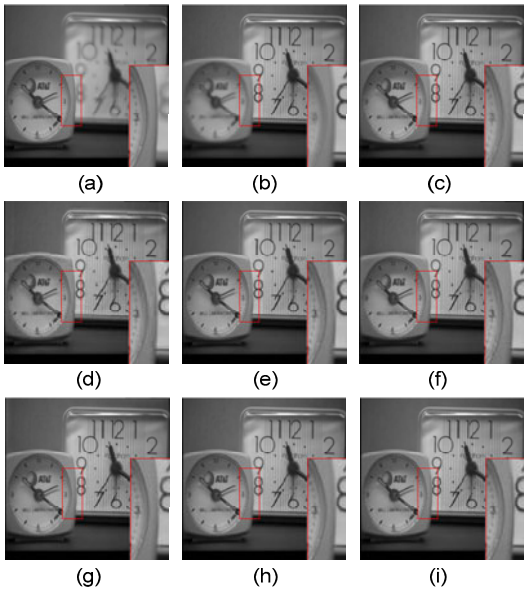

Fig. 7 The first set of gray images "Clock": (a) source image *A*; (b) source image *B*; (c) fused image by IMF; (d) fused image by DWT; (e) fused image by MWG; (f) fused image by NSCT; (g) fused image by GF; (h) fused image by CNN; (i) fused image by FD-FCN

References to color refer to the online version of this figure

Table 2 Quantitative evaluation of Fig. 6

Algorithm	MI	Q_{AB}^{F}	Q_e	Q_R	VIFF
IMF	8.064 029	0.732 390	0.726 478	1.098 585	0.599 546
DWT	5.726 883	0.617 876	0.718 810	0.926 814	0.442 699
MWG	7.741 579	0.729 557	0.727 447	1.094 336	0.598 777
NSCT	6.398 577	0.708 367	0.734 095	1.062 551	0.559 535
GF	7.707 244	0.733 896	0.728 883	1.100 844	0.601 925
CNN	8.104 806	0.735 962	0.728 961	1.103 943	0.605 657
FD-FCN	**8.182 621**	**0.745 621**	**0.738 705**	**1.156 895**	**0.606 512**

The best results are in bold

well in the letter's reflection on the desktop. In NSCT and GF, the effect of letter "P" on the left of images was not satisfying, and there were artifacts in some areas. Compared with previous images, although CNN performed well in all aspects, our algorithm performed better at the edge of the table. Table 4 shows the quantitative evaluation of "Peisi," which also indicates that our method outperformed others.

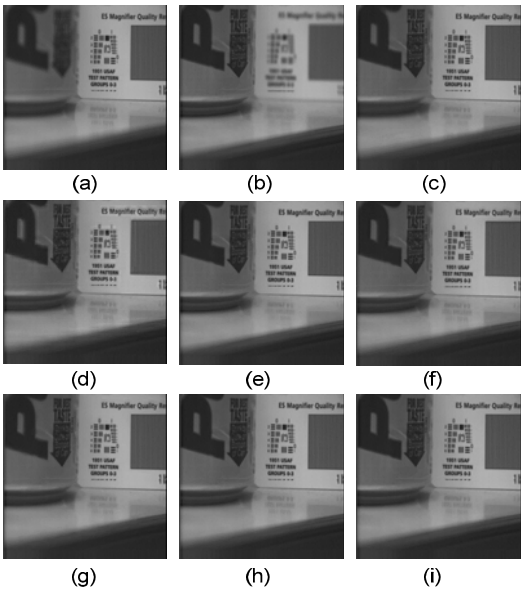

Fig. 8 The second set of gray images "Peisi": (a) source image *A*; (b) source image *B*; (c) fused image by IMF; (d) fused image by DWT; (e) fused image by MWG; (f) fused image by NSCT; (g) fused image by GF; (h) fused image by CNN; (i) fused image by FD-FCN

Finally, including the images shown in the study, we calculated the average of the objective evaluations in 20 sets of comparative experiments (Table 5). From the experimental results, we can see that our algorithm performed well.

To check whether there are significant differences between the proposed algorithm and others, we adopted Demšar's significance diagrams to show their performance on the five objective measures (Fig. 9). The diagram plots algorithms against mean ranks, whereby all methods are sorted according to their ranks. The line segment to the right of each algorithm represents its corresponding critical difference. That is, the right end of the line indicates from which mean rank onward another classifier is outperformed significantly. From the figure, it can be seen that the proposed algorithm is not significantly different from the CNN-based algorithm, but it performs significantly better than traditional ones.

4.3 Multi-focus image sequence fusion

The proposed algorithm FD-FCN can be easily extended to fuse an image sequence that contains more than two source images.

First, two source images were randomly selected for fusion, using the FD-FCN algorithm, and then the

Table 3 Quantitative evaluation of Fig. 7

Algorithm	MI	Q_{AB}^{F}	Q_e	Q_R	VIFF
IMF	6.906 191	**0.654 638**	0.781 511	0.971 937	0.572 523
DWT	6.157 212	0.622 260	0.784 093	0.933 390	0.498 242
MWG	6.919 281	0.650 379	0.793 032	0.978 569	0.567 951
NSCT	6.436 572	0.642 517	0.779 872	0.953 776	0.539 594
GF	6.623 621	0.648 170	0.781 085	0.972 255	0.559 643
CNN	6.914 834	0.653 912	0.791 296	**0.985 868**	0.569 685
FD-FCN	**6.931 562**	0.654 284	**0.795 830**	0.981 957	**0.569 727**

The best results are in bold

Table 4 Quantitative evaluation of Fig. 8

Algorithm	MI	Q_{AB}^{F}	Q_e	Q_R	VIFF
IMF	6.437 273	0.637 513	0.808 964	1.073 772	0.532 681
DWT	5.912 635	0.617 178	0.806 696	1.049 203	0.446 541
MWG	6.549 853	0.635 880	0.810 647	1.080 996	0.537 277
NSCT	6.336 545	0.646 272	0.814 889	1.098 662	0.528 496
GF	6.486 647	0.643 626	0.815 015	1.034 164	**0.545 344**
CNN	6.579 282	0.645 190	0.811 858	1.076 823	0.537 764
FD-FCN	**6.595 328**	**0.647 725**	**0.822 265**	**1.103 772**	0.540 985

The best results are in bold

Table 5 Quantitative evaluation of 20 sets of images

Algorithm	MI	Q_{AB}^{F}	Q_e	Q_R	VIFF
IMF	7.278 533	**0.712 536**	0.770 805	1.124 398	0.562 199
DWT	6.729 714	0.697 489	0.767 407	1.087 096	0.543 548
MWG	5.463 447	0.639 783	0.754 987	0.861 808	0.430 904
NSCT	7.198 559	0.706 364	0.764 045	1.099 248	0.549 624
GF	6.165 458	0.700 175	0.770 699	1.056 408	0.528 204
CNN	7.118 305	0.702 635	0.768 084	1.122 336	0.561 168
FD-FCN	**7.291 804**	0.712 533	**0.771 495**	**1.130 398**	**0.562 352**

The best results are in bold

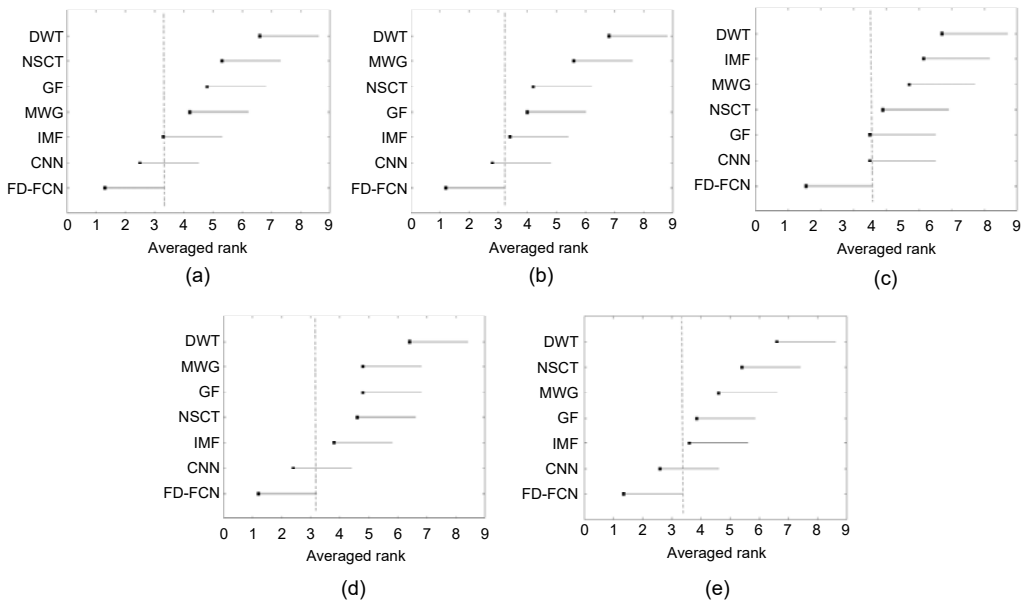

Fig. 9 Results of pairwise comparisons of all image fusion algorithms: (a) MI; (b) Q_{AB}^{F}; (c) Q_e; (d) Q_R; (e) VIFF

fused image was combined with another source image until all the source images were fused. Finally, we can obtain the fused images of the image sequence.

Figs. 10 and 11 show two sets of multi-focus image sequence fusion. There were 20 source images of size 720×480, and each image in the sequence had its own focus. In Fig. 10, we show one set of multi-focus sequence images, in which there was more detailed information, and the difference between focus and defocus in Fig. 11 was not evident.

4.4 Discussion

In the experiments, both subjective and objective evaluations have been conducted to compare the performance of the proposed algorithm with those of others. From the subjective experiments, it can be seen that our algorithm accurately resisted the influence of the dynamic area and selected focus area. In comparative experiments, the CNN-based and the proposed algorithms performed the best compared to traditional fusion algorithms. The reason is that they are trained using massive focused and defocused image patches, which causes them to easily make high-quality focus maps. To make the comparison results more reliable, we ran 20 groups of test images. The results showed that the proposed algorithm performed the best on the four measures MI, Q_e, Q_R, and VIFF. In summary, our algorithm has strong robustness.

However, there are two limitations in the proposed algorithm. The first is its long computation time, due to its many convolutional operations. Hence, the algorithm does not hold for highly real-time applications. The second is that FD-FCN tends to generate focus maps with isolated holes, so we need guided filtering to refine the maps. In future work, we

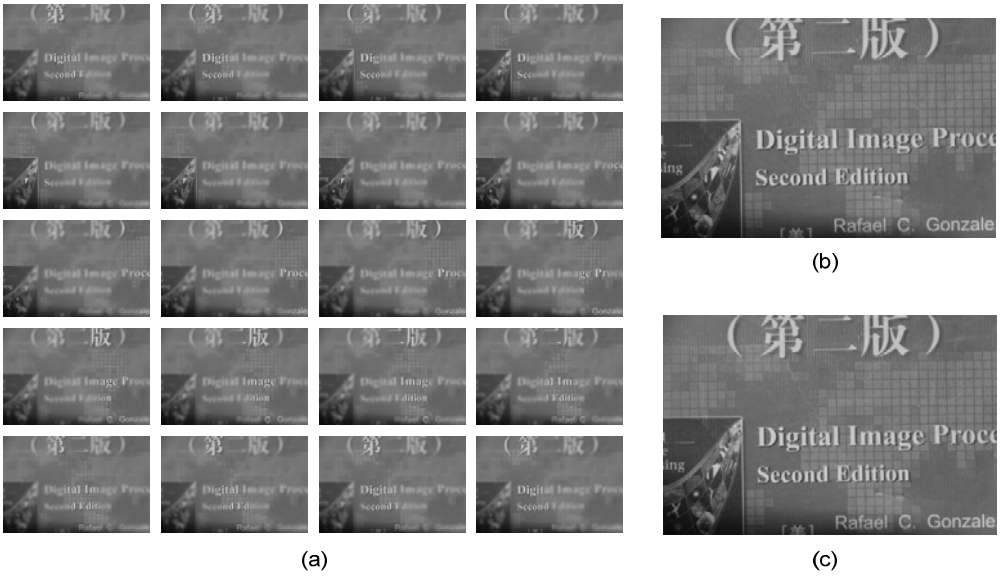

Fig. 10 The first set of multi-focus image sequence fusion "Cover": (a) source images; (b) fused image; (c) reference image

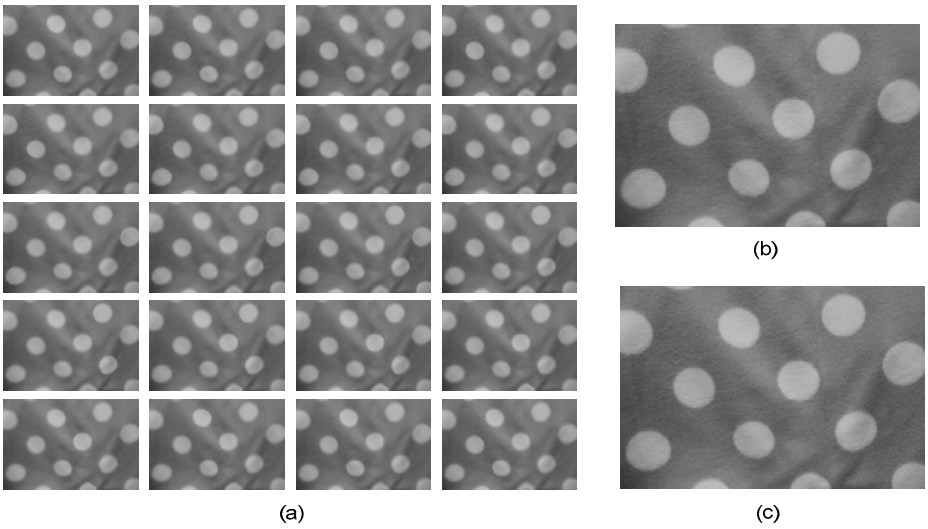

Fig. 11 The second set of multi-focus image sequence fusion "Circle": (a) source images; (b) fused image; (c) reference image

will resolve this problem by designing a multi-scale FD-FCN. Here, we discuss the threats to the validity of results from two perspectives:

1. Internal validity: In the experiments, all the performance evaluation measures and the codes of comparative algorithms were directly or indirectly derived from their authors. The parameter settings in these codes were the defaults. All of them were run on the same computer. Hence, the experimental settings guaranteed the fairness of the results.

2. External validity: We ran 20 groups of test images with various types. All of them were captured natural images rather than synthetic ones. Hence, they are applicable to other images.

5 Conclusions

In this paper, we have designed a fully convolutional network, FD-FCN, which can be used for

multi-focus image fusion. The dataset used for training the network is designed by Gaussian blur, and the process is easy to follow. The clarity of each pixel in the source images is determined by our designed network. Then, by comparing the clarity of the corresponding pixel, we can obtain a clear image. Combining the advantages of CNN with the reality of multi-focus image fusion, there are no problems of artificial threshold selection. We have conducted both subjective and objective evaluations to compare the performance of the proposed algorithm with those of six state-of-the-art algorithms. The quantitative evaluations of 20 sets of test images showed that the proposed algorithm is the best on the measures MI, Q_e, Q_R, and VIFF. In addition, Demšar's significance diagrams have been adopted to compare the performances of the fusion algorithms, and experiments showed that the proposed algorithm outperforms the five traditional algorithms.

In the future, we will design better-performing networks to obtain more accurate focus maps, or design a network that can directly generate higher-level results, such as direct output of fused images.

Contributors

Xiao-yu DONG designed the research. Rui GUO and Xuan-jing SHEN processed the data. Rui GUO drafted the manuscript. Xiao-li ZHANG helped organize the manuscript. Rui GUO and Xiao-li ZHANG revised and finalized the paper.

Compliance with ethics guidelines

Rui GUO, Xuan-jing SHEN, Xiao-yu DONG, and Xiao-li ZHANG declare that they have no conflict of interest.

References

Amin-Naji M, Aghagolzadeh A, Ezoji M, 2019. Ensemble of CNN for multi-focus image fusion. *Inform Fus*, 51:201-214. https://doi.org/10.1016/j.inffus.2019.02.003

Aslantas V, Kurban R, 2010. Fusion of multi-focus images using differential evolution algorithm. *Expert Syst Appl*, 37(12):8861-8870. https://doi.org/10.1016/j.eswa.2010.06.011

Ayyalasomayajula KR, Malmberg F, Brun A, 2019. PDNet: semantic segmentation integrated with a primal-dual network for document binarization. *Patt Recogn Lett*, 121:52-60. https://doi.org/10.1016/j.patrec.2018.05.011

Bai XZ, Zhang Y, Zhou FG, et al., 2015. Quadtree-based multi-focus image fusion using a weighted focus-measure. *Inform Fus*, 22:105-118. https://doi.org/10.1016/j.inffus.2014.05.003

Bavirisetti DP, Dhuli R, 2016. Fusion of infrared and visible sensor images based on anisotropic diffusion and Karhunen-Loeve transform. *IEEE Sens J*, 16(1):203-209. https://doi.org/10.1109/JSEN.2015.2478655

Chen Y, Qin Z, 2015. Gradient-based compressive image fusion. *Front Inform Technol Electron Eng*, 16(3):227-237. https://doi.org/10.1631/FITEE.1400217

Han Y, Cai YZ, Cao Y, et al., 2013. A new image fusion performance metric based on visual information fidelity. *Inform Fus*, 14(2):127-135. https://doi.org/10.1016/j.inffus.2011.08.002

He KM, Sun J, Tang XO, 2013. Guided image filtering. *IEEE Trans Patt Anal Mach Intell*, 35(6):1397-1409. https://doi.org/10.1109/TPAMI.2012.213

Huang W, Jing ZL, 2007. Multi-focus image fusion using pulse coupled neural network. *Patt Recogn Lett*, 28(9): 1123-1132. https://doi.org/10.1016/j.patrec.2007.01.013

Huhle B, Schairer T, Jenke P, et al., 2010. Fusion of range and color images for denoising and resolution enhancement with a non-local filter. *Comput Vis Image Understand*, 114(12):1336-1345. https://doi.org/10.1016/j.cviu.2009.11.004

Juočas L, Raudonis V, Maskeliūnas R, et al., 2019. Multi-focusing algorithm for microscopy imagery in assembly line using low-cost camera. *Int J Adv Manuf Technol*, 102(9-12):3217-3227. https://doi.org/10.1007/s00170-019-03407-9

Kim KI, Kwon Y, 2010. Single-image super-resolution using sparse regression and natural image prior. *IEEE Trans Patt Anal Mach Intell*, 32(6):1127-1133. https://doi.org/10.1109/TPAMI.2010.25

LeCun Y, Bengio Y, 1995. Convolutional networks for images, speech, and time-series. In: Arbib MA (Ed.), The Handbook of Brain Theory and Neural Networks. MIT Press, Cambridge, p.255-258.

Li ST, Kwok JT, Wang YN, 2002. Multifocus image fusion using artificial neural networks. *Patt Recogn Lett*, 23(8): 985-997. https://doi.org/10.1016/s0167-8655(02)00029-6

Li ST, Kang XD, Hu JW, 2013a. Image fusion with guided filtering. *IEEE Trans Image Process*, 22(7):2864-2875. https://doi.org/10.1109/TIP.2013.2244222

Li ST, Kang XD, Hu JW, et al., 2013b. Image matting for fusion of multi-focus images in dynamic scenes. *Inform Fus*, 14(2):147-162. https://doi.org/10.1016/j.inffus.2011.07.001

Li ST, Kang XD, Fang LY, et al., 2017. Pixel-level image fusion: a survey of the state of the art. *Inform Fus*, 33: 100-112. https://doi.org/10.1016/j.inffus.2016.05.004

Liu Y, Chen X, Peng H, et al., 2017. Multi-focus image fusion with a deep convolutional neural network. *Inform Fus*, 36:191-207. https://doi.org/10.1016/j.inffus.2016.12.001

Mustafa HT, Yang J, Zareapoor M, 2019. Multi-scale convolutional neural network for multi-focus image fusion. *Image Vis Comput*, 85:26-35. https://doi.org/10.1016/j.imavis.2019.03.001

Piella G, Heijmans H, 2003. A new quality metric for image fusion. Proc Int Conf on Image Processing, p.173-176. https://doi.org/10.1109/ICIP.2003.1247209

Qu GH, Zhang DL, Yan PF, 2002. Information measure for performance of image fusion. *Electron Lett*, 38(7):313-315. https://doi.org/10.1049/el:20020212

Ren WQ, Zhang JG, Xu XY, et al., 2019. Deep video dehazing with semantic segmentation. *IEEE Trans Image Process*, 28(4):1895-1908. https://doi.org/10.1109/TIP.2018.2876178

Saha A, Bhatnagar G, Wu QMJ, 2013. Mutual spectral residual approach for multifocus image fusion. *Dig Signal Process*, 23(4):1121-1135. https://doi.org/10.1016/j.dsp.2013.03.001

Saleem A, Beghdadi A, Boashash B, 2011. Image quality metrics based multifocus image fusion. 3rd European Workshop on Visual Information Processing, p.77-82. https://doi.org/10.1109/EuVIP.2011.6045547

Shelhamer E, Long J, Darrell T, 2017. Fully convolutional networks for semantic segmentation. *IEEE Trans Patt Anal Mach Intell*, 39(4):640-651. https://doi.org/10.1109/TPAMI.2016.2572683

Wang B, Yuan XY, Gao XB, et al., 2019. A hybrid level set with semantic shape constraint for object segmentation. *IEEE Trans Cybern*, 49(5):1558-1569. https://doi.org/10.1109/TCYB.2018.2799999

Xydeas CS, Petrovic V, 2000. Objective image fusion performance measure. *Electron Lett*, 36(4):308-309. https://doi.org/10.1049/el:20000267.

Yang Y, Huang SY, Gao JF, et al., 2014. Multi-focus image fusion using an effective discrete wavelet transform based algorithm. *Meas Sci Rev*, 14(2):102-108. https://doi.org/10.2478/msr-2014-0014

Yang Y, Tong S, Huang SY, et al., 2015. Multifocus image fusion based on NSCT and focused area detection. *IEEE Sens J*, 15(5):2824-2838. https://doi.org/10.1109/JSEN.2014.2380153

Zhang BH, Lu XQ, Jia WT, 2013. A multi-focus image fusion algorithm based on an improved dual-channel PCNN in NSCT domain. *Optik*, 124(20):4104-4109. https://doi.org/10.1016/j.ijleo.2012.12.032

Zhang Q, Levine MD, 2016. Robust multi-focus image fusion using multi-task sparse representation and spatial context. *IEEE Trans Image Process*, 25(5):2045-2058. https://doi.org/10.1109/TIP.2016.2524212

Zhao WD, Wang D, Lu HC, 2019. Multi-focus image fusion with a natural enhancement via a joint multi-level deeply supervised convolutional neural network. *IEEE Trans Circ Syst Video Technol*, 29(4):1102-1115. https://doi.org/10.1109/TCSVT.2018.2821177

Zhou ZQ, Li S, Wang B, 2014. Multi-scale weighted gradient-based fusion for multi-focus images. *Inform Fus*, 20:60-72. https://doi.org/10.1016/j.inffus.2013.11.005

Zhu XB, Zhang XM, Zhang XY, et al., 2019. A novel framework for semantic segmentation with generative adversarial network. *J Vis Commun Image Represent*, 58:532-543. https://doi.org/10.1016/j.jvcir.2018.11.020