Signal Processing: Image Communication 85 (2020) 115852

Contents lists available at ScienceDirect

Signal Processing: Image Communication

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

Perceptual quality assessment for multimodal medical image fusion

Lu Tang [a], Chuangeng Tian [c], Leida Li [d], Bo Hu [e], Wei Yu [e], Kai Xu [a,b,*]

[a] School of Medical Imaging, Xuzhou Medical University, China
[b] Department of Radiology, Affiliated Hospital of Xuzhou Medical University, China
[c] School of Information and Electrical Engineering, Xuzhou University of Technology, China
[d] School of Artificial Intelligence, Xidian University, China
[e] School of Information and Control Engineering, China University of Mining and Technology, China

ARTICLE INFO

Keywords:
Subjective and objective quality assessment
Multimodal medical image fusion
MMIF image database
Pulse coupled neural network
Non-subsampled contourlet transform

ABSTRACT

Recent years have witnessed that the multimodal medical image fusion (MMIF) plays critical roles in clinical diagnostics and treatment. Many MMIF algorithms have been proposed to improve the MMIF images quality. The quality of multimodal medical fused images will significantly affect the results of the clinical diagnosis. However, little work has been designed to evaluate the effectiveness of MMIF algorithms and the quality of MMIF images. To this end, this paper presents a perceptual quality assessment method for MMIF. A MMIF image database (MMIFID) is first built to employ the classical MMIF algorithms, and the subjective experiment is conducted to assess the quality of each fused image. Then, a no-reference objective method is proposed for the perceptual quality evaluation of MMIF images,which uses Pulse Coupled Neural Network (PCNN) in Non-subsampled Contourlet Transform (NSCT). A fused image is decomposed by NSCT into low frequency sub-band (LFS) and high frequency sub-band (HFS). It is used to motivate the PCNN processing, and large firing times are employed to measure LFS and HFS. Finally, two components evaluation results are combined to obtain the overall objective quality score. Experimental results based on the MMIFID indicate that our presented method outperforms the existing image fusion quality evaluation metrics, and it provides a satisfactory correlation with subjective scores, which shows effectiveness in the quality assessment of medical fused images.

1. Introduction

With the advent of computer technology, there are different imaging modalities in clinical use, such as ultrasound (US), computed tomography (CT), T1 T2 sequences of magnetic resonance imaging (MRI), positron emission tomography (PET) and single photon emission computed tomography (SPECT). Different imaging modalities supply respective specific information of human body, such as MRI give normal and pathological soft tissue,whereas PET provide blood flow and flood activity [1,2]. By fusing MRI and PET images, a single image that effectively provide morphologic anatomical structures and metabolic activities simultaneously can be obtained [3,4]. Hence, many multimodal medical image fusion (MMIF) schemes have become attracted wide attention. But the performance evaluation of these MMIF algorithms and fused images has not been fully understood. Although subjective evaluation is a direct and reliable way. It is extremely time-consuming. Furthermore, it is hard to be merged into the designing and optimization of image fusion. So, objective quality evaluation metrics attract much attention [5–8]. In the literature, some general-purpose no-reference (NR) quality measures have been proposed for fused images [9–15]. Among these quality metrics, they have achieved effectiveness in all kinds of image fusion types, including multi-exposure image fusion, multi-focus image fusion, and multimodal medical image fusion, they are not designed for MMIF, which show limited performance in evaluating the MMIF algorithms and the quality of medical fused images. So quality assessment metrics for MMIF specifically designed are eagerly desired.

As we know, ground-truth subjective score is very important for assessing the performance of any new objective strategy. Multimodal medical image fusion image database (MMIFID) can be directly employed for the test and comparison of the effectiveness and efficiency of different MMIF algorithms and fused images. So a MMIFID is firstly constructed. Established on the database, we conduct the subjective experiment to collect the subjective ratings. MMIF algorithms are then evaluated using the subjective study, which can evaluate and compare the quality of the fused images.

Based on the subjective database, a perceptual quality assessment metric is proposed for MMIF. As there are no perfect reference images in real-world medical imaging, and after many years of medical education and clinical experience, radiologists do not have any reference images for diagnosis. Thus, no reference metric is recommended for assessing medical fused images. In image processing, plenty of studies in

* Corresponding author.School of Medical Imaging, Xuzhou Medical University, China.
E-mail address: yingxiangxueyuan@163.com (K. Xu).

https://doi.org/10.1016/j.image.2020.115852
Received 23 April 2019; Received in revised form 23 December 2019; Accepted 7 April 2020
Available online 20 April 2020

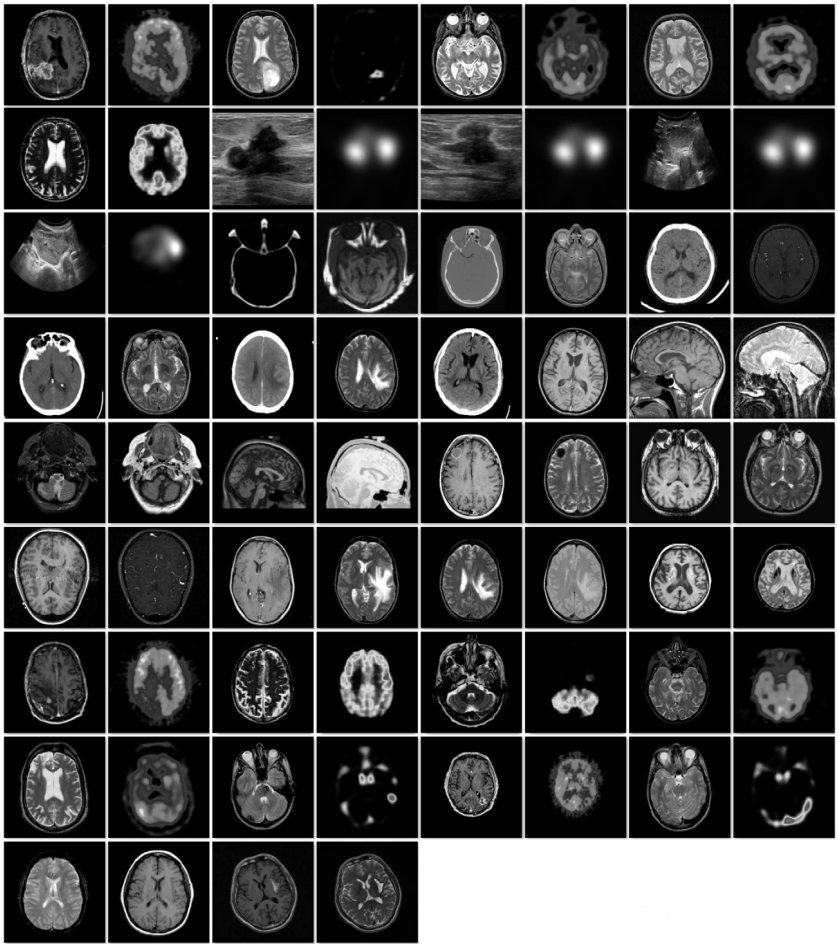

Fig. 1. 34 pairs of different modalities medical images used to build the MMIFID database.

the literature indicate that multiscale transform (MST)-based methods (e.g., nonsubsampled contourlet transform (NSCT) and nonsubsampled shearlet transform (NSST)) have exhibited significant advantages on account of their higher effectiveness in image representation. Many MMIF algorithms are introduced in a MST manner to pursue perceptually good results. In recent years, pulse coupled neural network (PCNN) under MST-based medical image fusion methods have been proposed. Here, we just give a few latest examples. Yin et al. [16] proposed a medical image fusion method based on NSCT and PCNN. Huang et al. [17] used NSCT and PCNN for the fusion of SPECT and CT images to improve the quality of fused brain images. Jin et al. [18] proposed simplified PCNN based on non-subsampled shearlet transform (NSST-PCNN) for multimodal sensor medical image fusion. Among these methods, PCNN is often used to extract the activity level of the decomposed coefficients obtained by a certain MST. The firing times of each output neuron over a number of iterations are typically employed to measure the activity level of its corresponding coefficient. Thus, PCNN based on MST methods have been verified a very suitable tool for processing medical image. In this article, we aimed to extract medical images features effectively for quality assessment. NSCT has better performance due to flexible multiscale, multidirectional, and shift-invariant image decomposition [19–21]. Pulse Coupled Neural Network (PCNN) is a bionic feedback neural network [22,23], which is a single-layer two-dimensional horizontally connected neuron array with a one to one correspondence between image pixels and neurons. A fused image represent with the different features. The features sensitive to human visual system (HVS) exist in different scales of images [24–26]. The NSCT coefficient value used to motivate PCNN can measure importance/contribution of fused image [27–29]. Based on the above advantages, NSCT and PCNN are first introduced for quality assessment of MMIF. First, the fused images are decomposed into low-frequency sub-band (LFS) and high-frequency sub-band (HFS) via NSCT, which is used to stimulate PCNN-processing. Then, large firing times are selected to measure LFS and HFS. Finally, two parts evaluation results are integrated to generate the final quality assessment score. The experimental results on the MMIFID database indicate that the presented scheme is better than the existing image fusion quality evaluation metrics.

The rest of this paper is organized as follows. In Section 2, subjective study is briefly introduced. Section 3 presents the theories of NSCT and PCNN, and provides a detailed analysis of proposed quality metric. Experimental results and analysis are presented in Section 4. Section 5 gives conclusion and discussion.

2. Subjective study of medical fused images

2.1. Database construction and subjective experiment

34 pairs of images (256 × 256) are used to construct the subjective database, which are shown in Fig. 1. These image pairs are selected to cover different imaging modalities, which are divided into CT and MRI, MRI-T1 and MRI-T2, B ultrasound and SPECT, MRI and PET, MRI and SPECT. Eight representative MMIF algorithms are selected, and 272 medical fused images are generated, where MMIF algorithms include medical image fusion with nuclear norm minimization [30] (NNM), image fusion with multi scale transform and sparse representation [15] (LP-SR), volumetric medical image fusion with cross-scale coefficient selection [31] (CSCS), image fusion with guided filtering [32] (GFF),

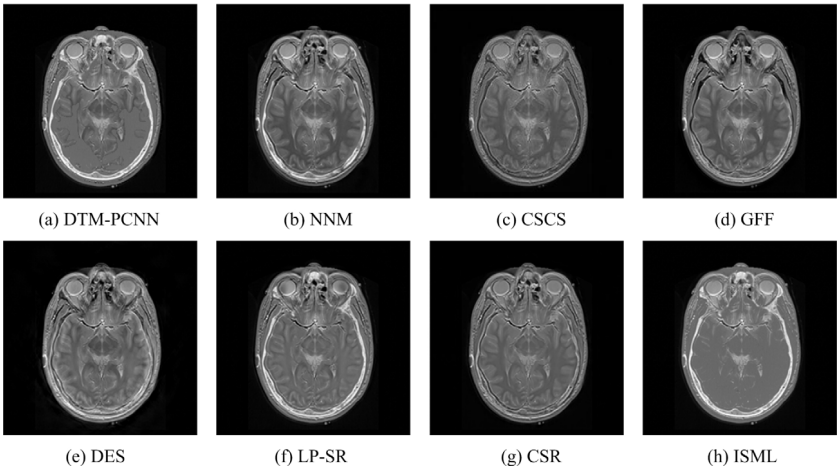

Fig. 2. An example of fused images (a)–(h) created by different MMIF algorithms..

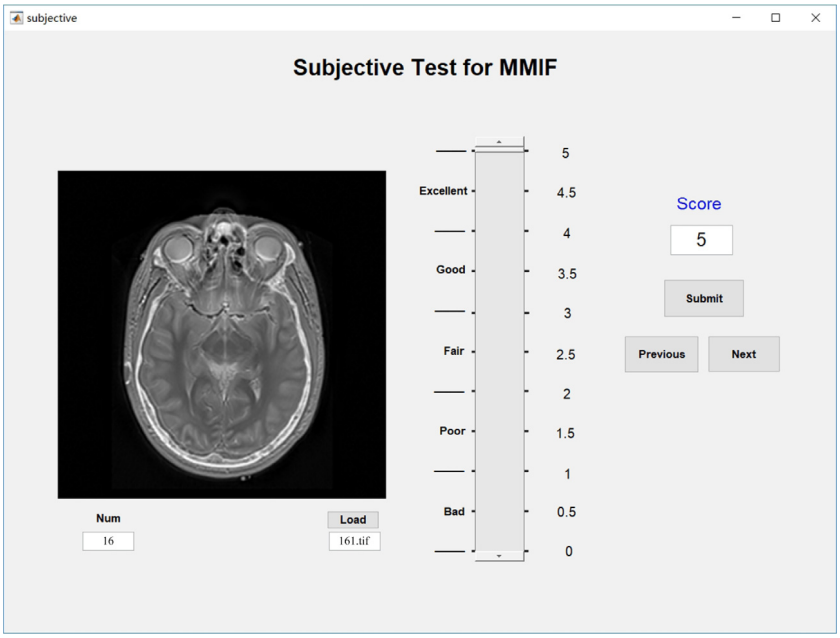

Fig. 3. GUI interface used for quality rating.

NSCT-based multimodal medical image fusion using pulse-coupled neural network and modified spatial frequency [33] (DES), medical image fusion with improved sum-modified-laplacian [34] (ISML), image fusion with convolutional sparse representation [35] (CSR), multimodal medical image fusion with discrete tchebichef moments and pulse coupled neural network [3] (DTM-PCNN). In DTM-PCNN, Tchebichef moments have been found effective in image analysis due to their superior capabilities of feature representation. Tchebichef moments are defined directly on image coordinates, so there is no approximation error in the computation of Tchebichef moments [3,36,37]. Fig. 2 gives an example with fusion results produced by eight fusion algorithms.

Different from natural image quality assessment, the quality evaluation of MMIF images is usually from radiologists ′ perspective, which is normally focused on the diagnostic task. So, 20 radiologists participated in the subjective test, all with medical education background of imaging diagnosis related researches. In order to conveniently obtain the image quality, a custom MATLAB figure window is used as to make the images on the screen, which is shown in Fig. 3. The fused images were arranged randomly to avoid any position selection bias. They can observe all the necessary details under normal light condition by adopting a LCD (1920 × 1280) display, similar to those applied during clinical routine. In fact, the subjective visual quality evaluation depends on the diagnostic task that refers to radiologists give a confidence rating concerning the presence of a lesion. In other word, it depends on the information that can be extracted from the fused images themselves. Since the ultimate goal of MMIF is to facilitate the diagnostic process, any visual artifact in the fused images may obstruct diagnostic conclusions and lead to severe consequences. Radiologists would prefer high quality images for accurate diagnose, a high quality fused image will increase the confidence in diagnosing and subsequently lead to the right treatment. Thus, they give a ranking score 0–5 within a continuous range to each medical fused image, where 0–1 means that image fusion performance is bad, it is very annoying and seriously affects the diagnosis. 1–2 indicates fusion result is poor, which is annoying and affects the diagnosis. 2–3 means the image fusion quality is fair, it is slightly annoying and slightly affects the diagnosis. 3–4 means the image fusion result is good, it is perceptible, but not affects the diagnosis. 4–5 indicates that fusion performance is excellent, it is imperceptible and accurate diagnosis. The different subjective quality score provide us to easily assess the performance of objective metrics.

Table 1
SRCC, KRCC, PLCC and RMSE values of average subject.

Image set	SRCC		KRCC		PLCC		RMSE	
	$\mu(subject)$	$\sigma(subject)$	$\mu(subject)$	$\sigma(subject)$	$\mu(subject)$	$\sigma(subject)$	$\mu(subject)$	$\sigma(subject)$
Average	0.87	0.07	0.78	0.10	0.89	0.07	0.16	0.05

Table 2
K-S test result of individual MMIF algorithms.

K-S test	DTM-PCNN	NNM	LP-SR	CSCS	GFF	DES	ISML	CSR
DTM-PCNN	0	1	1	1	1	1	−1	1
NNM	−1	0	−1	1	−1	−1	−1	1
LP-S	−1	1	0	1	1	1	−1	1
CSCS	−1	−1	−1	0	−1	−1	−1	−1
GFF	−1	1	−1	1	0	0	−1	1
DES	−1	1	−1	1	0	0	−1	1
ISML	1	1	1	1	1	1	0	1
CSR	−1	−1	−1	1	−1	−1	−1	0

2.2. Subjective data analysis

Generally, to calculate more accurate evaluation results, the abnormal values are discarded using the outlier removal method. However, there are any abnormal values, maybe after years of medical education and clinical experience, radiologists diagnose without any abnormal values. After the subjective test, we can obtain 20 subjective scores for each fused image. The subjective scores of each fused image is averaged to obtain the final quality score, i.e. the mean opinion score (MOS) [38,39]:

$$\bar{u} = \frac{1}{N}\sum_{i=1}^{N} u_i \tag{1}$$

where u_i denotes the subjective score of the i_{th} radiological doctors, N denote the 20 radiologists. Then the standard deviation of the subjective scores is computed [38,39]:

$$\sigma = \sqrt{\sum_{i=1}^{N} \frac{(u_i - \bar{u})^2}{(N-1)}} \tag{2}$$

For performance evaluation, spearman rank order correlation coefficient (SRCC), kendall rank order correlation coefficient (KRCC), pearson linear correlation coefficient (PLCC) and root -mean-square error (RMSE) are adopted as the evaluation criteria [40–42]. Specifically, SRCC and KRCC are two most popular rank correlation indicators, which can predictive consistency between MOS values and individual subject ratings, while PLCC and RMSE are used for measuring prediction accuracy. For each image set, we calculate the SRCC, KRCC, PLCC and RMSE values between MOS values and individual subject ratings. When this is conducted for all 20 subjects, we calculate the mean (μ) and standard deviation (σ) of SRCC, KRCC, PLCC and RMSE values over all subjects. The average performance cross all individual subjects scores over all 34 image set are showed in Table 1. From Table 1, we can see that average subject scores of SRCC, KRCC, PLCC and RMSE perform quite consistently with relatively low variations for different image modalities.

2.3. Performance of existing MMIF algorithms

In practice, with many MMIF schemes at present, no single MMIF algorithm performs best for all images, and this is not easy to choose the method with best performance for image fusion. Fortunately, since the subjective test results have been obtained, relative performance gradings of the MMIF methods can be measured by implementing a systematic evaluation of MMIF algorithms. As we know, a high quality fused image will increase the confidence in diagnosing and subsequently lead to the right treatment, a better MMIF algorithm may obtain higher quality fused images for accurate diagnosis, which will have higher scores by radiologist. As a result, we can use the average MOS values to assess the relative performances of different MMIF algorithms. Fig. 4 shows the average MOS values and the corresponding mean and standard deviation for fused images, which are generated by the eight MMIF algorithms over all 34 image sets. It is observed from Fig. 4 that ISML achieves the best performance on average, this indicates that the fusion algorithm performs better than other algorithms. LP-SR is the second best on average. DTM-PCNN produces quite similar average performance and larger standard deviations than LP-SR. This suggests that DTM-PCNN might perform well on some images but perform bad on some other images, thus the overall performance cannot be guaranteed. The average MOS value of CSCS is the lowest, which indicates that its performance is the worst.

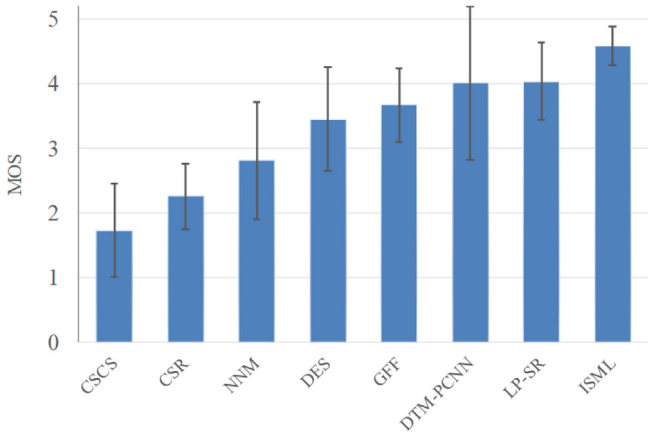

Fig. 4. Mean and standard deviation of subjective rankings of individual MMIF algorithms across all image sets.

In order to demonstrate performance of the MOS values in different fusion algorithm, we compare the 8 pairs of MMIF algorithm by K-S test method. Considering that the same radiologist scored fused images for each fusion algorithm, we used paired samples K-S tests to compare in the experiment. Our data is divided into multiple subsets according to different fusion methods, and each subset performs only one statistical method. Table 2 shows the result of K-S test of individual MMIF algorithms, where a symbol 0 denotes the null hypothesis that the data in two algorithms are from the same continuous distribution, and cannot be refused at the alpha level. A symbol 1 denotes the row algorithm is better than the column algorithm, whereas a symbol -1 denotes the row algorithm is inferior to the column algorithm. From Table 2, we can see that there is no significant difference between the means in a pairs fusion algorithms (GFF and DES), and their fusion effect has little difference. The results of others pair of fusion algorithms show that their fusion effects can be considered as differences.

3. Objective quality assessment of MMIF images

3.1. Nonsubsampled contourlet transform

Medical images typically consist of smooth regions and edges, and they can be well characterized by their features. NSCT is a image decomposition transform tool with shift-invariant, multiscale and multidirection, which can be utilized to process different features [19–21]. It is based on two filter bank (FB), the nonsubsampled pyramid FB

(NSPFB) and the nonsubsampled directional FB (NSDFB). The NSPFB performs multiscale decomposition, which divides an image into a low frequency sub-band and a high frequency sub-band. With k decomposition layers, NSPFB generates $k+1$ sub-images, including one low frequency image and k high frequency images, which have the same size as source image. NSDFB provides direction decomposition, which decomposes the high frequency sub-bands image from NSPFB in each layer, producing 2^l directional sub-images, and their size is the same to the source image, where l is the number of decomposition directions. So, NSDFB gives the NSCT with more accurate directional detail information. The multi-scale and multi-directional characteristics of NSCT are very useful in extracting more directional information.

3.2. Pulse coupled neural network

PCNN is a based cat visual cortex biologically inspired spiking neural network, which is no training processing. The output pulses of PCNN can describe the regional and detailed information of an input image effectively. Because of its superior performance, PCNN has been widely used in medical image processing. The simplified PCNN model is given in Fig. 5. There are three modules: the dendritic (feeding input $E_{\upsilon,v}$ and linking input $I_{\upsilon,v}(n)$), the linking modulation $M_{\upsilon,v}(n)$ and the pulse generator $F_{\upsilon,v}(n)$ [22,23], which are denoted by:

$$E_{\upsilon,v}(n) = S_{\upsilon,v} \tag{3}$$

$$I_{\upsilon,v}(n) = e^{-\alpha_L} I_{\upsilon,v}(n) + V_L \sum_{k,l} W_{\upsilon,v,k,l} F_{\upsilon,v}(n-1) \tag{4}$$

$$M_{\upsilon,v}(n) = E_{\upsilon,v}(n)\left[1 + \beta I_{\upsilon,v}(n)\right] \tag{5}$$

$$T_{\upsilon,v}(n) = e^{-\alpha_T} T_{\upsilon,v}(n-1) + V_T F_{\upsilon,v}(n-1) \tag{6}$$

$$F_{\upsilon,v}(n) = \begin{cases} 1, M_{\upsilon,v}(n) > T_{\upsilon,v}(n) \\ 0, M_{\upsilon,v}(n) \le T_{\upsilon,v}(n) \end{cases} \tag{7}$$

where W is the synaptic weight matrices υ, v represent the pixel locations k, l represent the dislocation in a symmetric neighborhood surrounding a pixel. $S_{\upsilon,v}(n)$ stands for the external stimulus. In this paper, gray value of the pixel instead of external stimulus of PCNN. V_L and α_L represent normalizing constants. β is the linking parameter, which denotes the weight of linking field. α_T and V_T denote attenuation coefficient and threshold magnitude coefficient, respectively.

3.3. Proposed quality metric

In our work, a quality evaluation metric of MMIF image is proposed via PCNN in NSCT. Fig. 6 shows the flowchart of the proposed metric. Uniformly, let F represents the fused image. The proposed fusion method is stated in following steps.

Firstly, the fused image F is transformed by NSCT to obtain one low-frequency sub-band $\left\{F_\iota^L\right\}$ and a series of high-frequency sub-band $\left\{F_{\iota,\theta}^H\right\}$ at each level $\iota \in [1, n]$ and direction θ.

Secondly, the $\left\{F_\iota^L\right\}$ and $\left\{F_{\iota,\theta}^H\right\}$ are used as the stimulus of the PCNN, respectively.

$$T_\iota^L = PCNN\left(F_\iota^L\right), T_{\iota\theta}^H = PCNN\left(F_{\iota\theta}^H\right) \tag{8}$$

where $PCNN(\cdot)$ denotes the PCNN functions by Eqs. (3)–(7), which described in Section 3.2. The firing times matrix $\left\{T_\iota^L\right\}$ and $\left\{T_{\iota,\theta}^H\right\}$ denote the total fired times motivated by $\left\{F_\iota^L\right\}$ and $\left\{F_{\iota,\theta}^H\right\}$, respectively. If the iteration number $n = N_{max}$, N_{max} is the max iteration times, then iteration stops. The parameters of PCNN used in the experiments are set as follows: $W = [0.7071, 1, 0.7071; 1, 0, 1; 0.7071, 1, 0.7071]$, $V_L = 1.0$, $\alpha_L = 0.0693$, $\alpha_T = 0.2$, $N_{max} = 200$.

Next, the quality of low-frequency coefficient $\left\{T_\iota^L\right\}$ at location (m, n) is given by:

$$Q_L = T_\iota^L(m, n) \tag{9}$$

where $m \in \{1, 2, \ldots, M\}, n \in \{1, 2, \ldots, N\}$, $T_\iota^L(m, n)$ is firing times matrix of the low-frequency sub-band coefficient at location (m, n). Meanwhile, the sum of firing times value is adopted to detect the quality of high-frequency sub-band, which is computed and denoted by:

$$Q_H = \sum_{m}^{M} \sum_{n}^{N} T_{\iota,\theta}^H(m, n) \tag{10}$$

Finally, the evaluation results of two components are merged to get the overall quality score:

$$Q = a\left(Q_L\right)^b + (1-a)\left(Q_H\right)^c \tag{11}$$

where a, b and c are parameters, which are utilized to balance the relative relationship of the two components in the quality evaluation of MMIF images. Parameters are determined by experiments. In this paper, we set $a = 0.8, b = 2.5, c = 3$.

4. Experimental results

4.1. Performance evaluation

To verify the effectiveness of the proposed quality evaluation metric, we combine Q_L and Q_H into the MMIF quality metrics with Eq. (11) and their performances on the MMIFID database is evaluated. Table 3 summarizes the experimental results together with the performances of each component $\left(Q_L, Q_H\right)$ and proposed metric that combines the two components in accordance with SRCC, KRCC, PLCC and RMSE. We have computed the results of each image set, and the average performance over all 34 image set are shown in Table 3. It is observed from Table 3 that our metric is very valid for MMIF images, which indicates that both two components are only provide a moderate correlations with subjective opinions, after combining these two components, the proposed metric obtains performance improvement on two parts.

To further check the validity of the proposed quality model, we also make comparison with eleven state-of-the-art image fusion quality metrics: (1) Gradient based fusion metric Q_G [9], which evaluates the success of gradient information from source images is injected in to the fused image from the source images. (2) Structure based metrics Q_S [10], which utilizes the structure similarity to measure the structural information in the fused image. (3) Edge information Q_E [11], which evaluates the amount of edge information transmitted from the source image to the fused image. (4) Phase congruency Q_P [12], which reflects image salient features. (5) Ratio of spatial frequency error based metrics Q_{RSFE} [13], which addresses perceptual assessment with sorting in the fused image. (6) Mutual information based metrics Q_{MI} [14], which measures correlation between the two input images, or how much the information of the source image is kept in the fused image. (7) Entropy Q_{EN} [15] measures the amount of information in the fused image. (8) Optimizing structural similarity index Q_{OSSI} [43] evaluates fused image. (9) In-depth analysis of Tsallis entropy-based metrics Q_{TEIA} [44] measures fused image quality. (10) Q_{BMPRI} [45] computes the distorted image quality. (11) Multi-task end-to-end optimized deep neural network Q_{MEON} [46], which evaluates blind image quality. In the above strategies, the larger the values of Q_G, Q_S, Q_E, Q_P, Q_{MI}, Q_{EN}, Q_{OSSI}, Q_{TEIA}, Q_{BMPRI}, Q_{MEON} denote better quality, however, the smaller the values of Q_{RSFE} shows better quality. We calculate the SRCC, KRCC, PLCC and RMSE values of all 34 image sets for twelve image fusion quality metrics (including the eleven existing models described above and proposed metric). The average performances over all image sets are summarized in Table 4. It is observed from Table 4 that our proposed strategy obtains the best performance on the average values of SRCC, KRCC, PLCC and RMSE. Unfortunately, the eleven state-of-the-art image fusion quality evaluation metrics seem be impossible to provide sufficient quality to predict MMIF images.

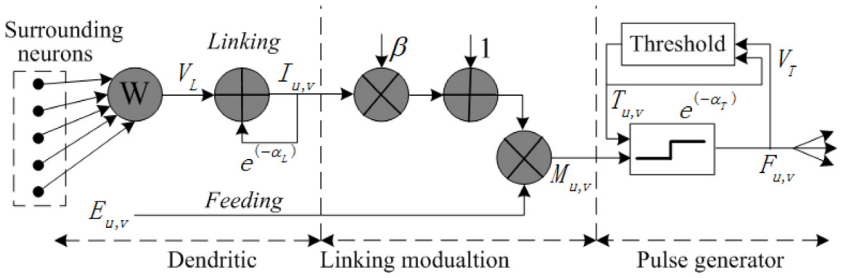

Fig. 5. Diagrammatic sketch of simplified PCNN.

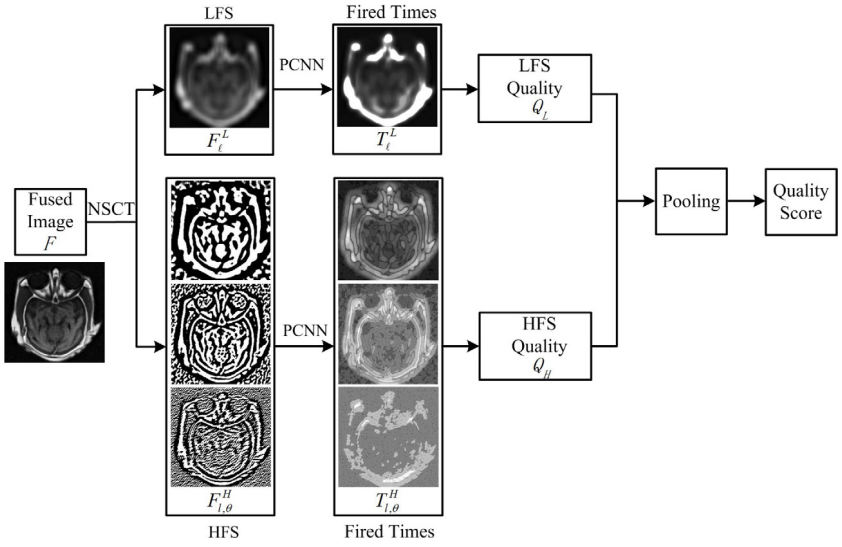

Fig. 6. Flowchart of the proposed quality metric.

Table 3
SRCC, KRCC, PLCC and RMSE performance of proposed model and its two components.

Image set	SRCC			KRCC			PLCC			RMSE		
	Q_L	Q_H	Pooling	Q_L	Q_H	Pooling	Q_L	Q_H	Pooling	Q_L	Q_H	Pooling
Average	0.71	0.46	0.73	0.58	0.35	0.61	0.75	0.60	0.79	0.29	0.30	0.27

Table 4
SRCC, KRCC, PLCC and RMSE performance assessment of eleven existing schemes and the proposed model on MMIFID database.

Metric	SRCC	KRCC	PLCC	RMSE
Q_G	0.61	0.47	0.45	0.36
Q_S	0.67	0.52	0.65	0.28
Q_E	0.54	0.42	0.52	0.40
Q_P	0.33	0.29	0.33	0.41
Q_{RSFE}	0.43	0.34	0.50	0.33
Q_{MI}	0.55	0.44	0.51	0.39
Q_{EN}	0.39	0.28	0.33	0.50
Q_{OSSI}	0.31	0.23	0.20	0.44
Q_{TEIA}	0.42	0.34	0.16	0.43
Q_{BMPRI}	0.25	0.21	0.28	0.55
Q_{MEON}	0.26	0.23	0.27	0.39
Proposed	**0.73**	**0.61**	**0.79**	**0.27**

For better understanding the relationships between the above MMIF quality measures, a dendrogram plot is used to represent the similarities between the MMIF quality metrics. To be specific, firstly, the KRCC values between the ten MMIF quality metrics are calculated and averaged for image sets. Then, KRCC values are converted to distances by using the dendrogram, and Euclidean distance is adopted as a distance metric, to cluster KRCC values, the arithmetic average algorithm of unweighted pair group measure are used [39], which are shown in Fig. 7. The horizontal axis is used as the leaf nodes, which shows the MMIF quality metric and MOS. The vertical axis represents distances between the nodes. Node height is the distance between the left and right sub-branch clusters. From the evaluation results in Fig. 7, it is clear that our metric produces the best results among all the existing algorithms, meanwhile it is close to MOS value.

4.2. Performance ranking consistent

In MMIFID database, there are two image groups, Group 1 (G1): fused images generated by fusion of anatomical images and functional images; Group 2 (G2): fused images generated by fusion of anatomical images and anatomical images. For a specific MMIF algorithm, the average MOS value is utilized to measure its performance, which is used as ground truth. The average objective score is also calculated. The rankings consistent performance between the subjective rankings and objective rankings is listed in Table 5, where ranking 1 shows the best performance and 8 represents the worst performance, and inconsistent rankings are marked in boldface. By contrast, the proposed method is less inconsistent rankings than other quality metric, it demonstrates the best performance.

Table 5
Ranking of the MMIF algorithms according to the subjective scores and predicted scores obtained by twelve objective metrics in two groups.

Group	MMIF	MOS	Q_G	Q_S	Q_E	Q_P	Q_{RSFE}	Q_{MI}	Q_{EN}	Q_{OSSI}	Q_{TEIA}	Q_{BMPRI}	Q_{MEON}	Proposed
G1	IMSL	1	1	1	1	**5**	**6**	1	**6**	1	1	**3**	**5**	1
	DTM-PCNN	2	2	2	2	**6**	**4**	2	**8**	**5**	2	**1**	2	2
	LP-SR	3	**5**	**5**	**5**	**4**	**2**	**5**	**5**	**2**	**7**	**4**	**4**	3
	GFF	4	**3**	**3**	4	**1**	**3**	4	**3**	**7**	**5**	**5**	**3**	**6**
	DES	5	**8**	**6**	**7**	**8**	5	**8**	**1**	**4**	**8**	**2**	**6**	5
	NNM	6	**4**	**4**	**3**	**3**	**1**	**3**	**4**	6	**3**	**8**	**1**	**4**
	CSR	7	**6**	7	**8**	**2**	7	7	7	**3**	**6**	7	7	7
	CSCS	8	**7**	8	**6**	**7**	8	**6**	**2**	8	**4**	**6**	8	8
G2	IMSL	1	1	1	1	**2**	**5**	1	**8**	**7**	1	1	**2**	1
	LP-SR	2	**5**	**5**	**4**	**4**	**3**	**4**	**4**	**6**	**7**	**4**	**6**	2
	DTM-PCNN	3	**2**	**2**	**2**	**5**	**1**	**2**	**7**	**8**	3	**5**	**1**	3
	GFF	4	4	**3**	**5**	**1**	**6**	**5**	**2**	**1**	4	**2**	**5**	**5**
	DES	5	**8**	**7**	**7**	**6**	**4**	**7**	**1**	**4**	**6**	**3**	**4**	**6**
	NNM	6	**3**	**4**	**3**	**7**	**2**	**3**	**3**	**3**	**2**	**8**	**3**	**4**
	CSR	7	**6**	**6**	**8**	**3**	7	**8**	**6**	**5**	**8**	**6**	7	7
	CSCS	8	**7**	8	**6**	8	8	**6**	**5**	**2**	**5**	**7**	8	8

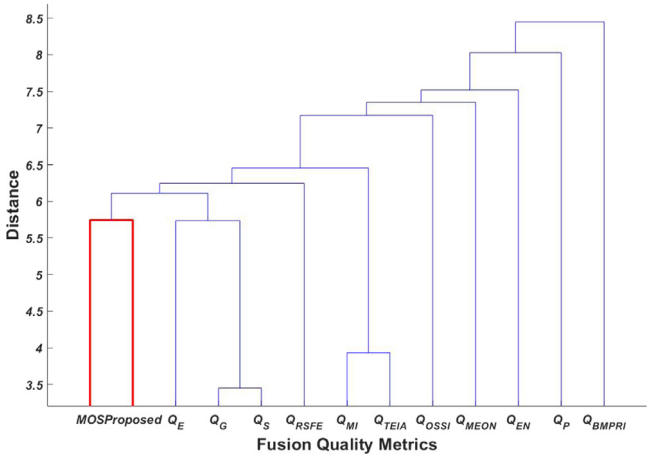

Fig. 7. Similarity dendrogram of MOS and objective image fusion quality measures.

4.3. Computational complexity

We also compare the performance of execution time for MMIF quality models. Specifically, the average time for predicting the scores of MMIF images with resolution 256 × 256 in MMIFID database is calculated to measure the computation complexity. The average execution time of twelve quality metrics are listed in Table 6. It is observed from Table 6 that most of the metrics are fast to calculate. The proposed are more computationally demanding. This mainly attributes to the decomposition operation and more PCNN iteration required. In implementation process, it takes up most of the time, but it demonstrates the best performance. In the future, we will try to explore other technologies to improve computational efficiency.

5. Conclusion and discussion

Multimodal medical image fusion has been extensively researched for improving clinical diagnostics and treatment. However, existing quality assessment methods are general-purpose image fusion quality measures, which are not designed for MMIF algorithms and MMIF images, and little work has been devoted to the quality assessment of MMIF images. This paper proposed a perceptual quality evaluation scheme for MMIF images. Firstly, we create a MMIF image database and conducted subjective experiments to obtain the ground truth of each MMIF image. Secondly, we present an objective quality assessment algorithm for MMIF via combining PCNN and NSCT. Finally, the performances of MMIF algorithms are compared based on MMIFID, and the effectiveness is verified. Experimental results show that our metric exhibits the highest consistency with the subjective results, and outperforms the existing state-of-the-art metrics, which is more applicable to evaluate MMIF images.

Table 6
Average execution time for predicting the quality scores in MMIFID database by different image quality metrics.

Metric	Time (s)	Metric	Time (s)
Q_G	0.0260	Q_{EN}	0.0007
Q_S	0.0148	Q_{OSSI}	1.0677
Q_E	0.0122	Q_{TEIA}	0.0092
Q_P	0.7945	Q_{BMPRI}	0.6178
Q_{RSFE}	0.0096	Q_{MEON}	0.0132
Q_{MI}	0.0054	Proposed	8.1631

As one of the first attempts on the quality assessment of MMIF images, our metric outperforms the existing ones, but it is still not so good (correlation coefficient < 0.8), thus, the proposed method has several limitations that need to be resolved or improved in the future research. Firstly, the proposed method is mainly based on the NSCT and PCNN, other technologies that have been successfully used in image quality assessment may be exploited in the context of MMIF, such as extracting image spatial features of structure and texture. Secondly, due to functional images (e.g., PET and SPECT) are color images. Appropriate consideration of color features may improve the performance of image fusion quality assessment models. Thirdly, the existing image fusion subjective databases are nearly 2-dimensional scenes, including our MMIF subjective database. But most medical images we encounter in clinical practice are often 3-dimensional scenes. It is useful to extend the proposed model to consider 3-dimensional scenes. Fourth, how to use quality assessment model as new optimization goal to redesign MMIF algorithm for obtaining better fused image quality is challenging and interesting problems yet to be explored. Finally, modern medical imaging scan post-processing platform often display fused image generated by merging different imaging modalities before diagnosis. Quality assessment and optimization methods may be applied to these extended medical applications.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (81771904), Project Funded by China Post doctoral Science Foundation (2018M642325). The Xu Zhou Science and technology Program, China (KC19146).

References

[1] J.N. Li, Y. He, W. Dong, L.J. Zhang, et al., Comparison of cardiac MRI with PET for assessment of myocardial viability in patients with coronary chronic total occlusion, Clin. Radiol. 74 (5) (2019) 410.e1–410.e9.
[2] L. Fernandez-Friera, V. Fuster, B. Lopez-Melgar, Vascular inflammation in subclinical atherosclerosis detected by hybrid PET/MRI, J. Am. Coll. Cardiol. 73 (12) (2019) 1371–1382.
[3] L. Tang, J.S. Qian, L.D. Li, J.F. Hu, X. Wu, Multimodal medical image fusion based on discrete tchebichef moments and pulse coupled neural network, Int. J. Imaging Syst. Technol. 27 (1) (2017) 57–65.
[4] J.S. Qian, B. Rong, W. Shen, J.F. Hu, L. Tang, Perceptual medical image fusion with internal generative mechanism, Electron. Lett. 53 (17) (2017) 1184–1186.
[5] S.Q. Wang, K. Gu, X.F. Zhang, W.S. Lin, S.W. Ma, W. Gao, Reduced-reference quality assessment of screen content images, IEEE Trans. Circuits Syst. Video Technol. 28 (1) (2018) 1–14.
[6] S.Q. Wang, K. Gu, X. Zhang, W.S. Lin, L. Zhang, S.W. Ma, W. Gao, Subjective and objective quality assessment of compressed screen content images, IEEE J. Emerg. Sel. Top. Circuits Syst. 6 (4) (2016) 532–543.
[7] K. Gu, J.F. Qiao, X.K. Min, G.H. Yue, W.S. Lin, D. Thalmann, Evaluating quality of screen content images via structural variation analysis, IEEE Trans. Vis. Comput. Graphics 24 (10) (2018) 2689–2701.
[8] V. Jakhetiya, K. Gu, W. Lin, Q. Li, S. Jaiswal, A prediction backed model for quality assessment of screen content and 3D synthesized images, IEEE Trans. Ind. Inf. 14 (2) (2018) 652–660.
[9] C.S. Xydeas, V.S. Petrovic, Objective image fusion performance measure, Electron. Lett. 36 (4) (2000) 308–309.
[10] C. Yang, J.Q. Zhang, X.R. Wang, A novel similarity based quality metric for image fusion, Inf. Fusion 9 (2) (2008) 156–160.
[11] S. Li, J.T. Kwok, Y. Wang, Combination of images with diverse focuses using the spatial frequency, Inf. Fusion 2 (3) (2001) 169–176.
[12] J. Zhao, R. Laganiere, Z. Liu, Performance assessment of combinative pixel-level image fusion based on an absolute feature measurement, Int. J. Innovative Comput. Inf. Control 3 (6) (2007) 1433–1447.
[13] Y. Zheng, E.A. Essock, B.C. Hansen, A.M. Haun, A new metric based on extended spatial frequency and its application to DWT based fusion algorithms, Inf. Fusion 8 (2) (2007) 177–192.
[14] G. Qu, D. Zhang, P. Yan, Information measure for performance of image fusion, Electron. Lett. 38 (7) (2002) 313–315.
[15] Y. Liu, S. Liu, Z. Wang, A general framework for image fusion based on multi-scale transform and sparse representation, Inf. Fusion 24 (2015) 147–164.
[16] M. Yin, X.N. Liu, Y. Liu, X. Chen, Medical image fusion with parameter-adaptive pulse coupled neural network in nonsubsampled shearlet transform domain, IEEE Trans. Instrum. Meas. 68 (1) (2019) 49–64.
[17] C.X. Huang, G.X. Tian, Y.S. Lan, Y.H. Peng, E.Y.K. Ng, Y.T. Hao, Y.Q. Cheng, W.L. Che, A new pulse coupled neural network (PCNN) for brain medical image fusion empowered by shuffled frog leaping algorithm, Front. Neurosci. 13 (2019) 1–10.
[18] X. Jin, G. Chen, J.Y. Hou, Q. Jiang, D. Zhou, S. Yao, Multimodal sensor medical image fusion based on nonsubsampled shearlet transform and S-PCNNs in HSV Space, Signal Process. 153 (2018) 379–395.
[19] A.L. Cunha, J.P. Zhou, M.N. Do, The non-subsampled contourlet transform: theory, design and applications, IEEE Trans. Image Process. 15 (10) (2006) 3089–3101.
[20] G. Bhatnagar, Q.M.J. Wu, Z. Liu, Directive contrast based multimodal medical image fusion in NSCT domain, IEEE Trans. Multimed. 15 (5) (2013) 1014–1024.
[21] Y. Liu, S. Liu, Z. Wang, Medical image fusion by combining nonsubsampled contourlet transform and sparse representation, Pattern Recognit. 484 (2014) 372–381.
[22] X. Xu, G. Wang, S. Ding, Y. Cheng, X. Wang, Pulse-coupled neural networks and parameter optimization methods, Neural Comput. Appl. 4 (2016) 1–11.
[23] Z. Wang, Y. Ma, F. Cheng, L. Yang, Review of pulse-coupled neural networks, Image Vis. Comput. 28 (1) (2010) 5–13.
[24] Y. Yang, S. Huang, J. Gao, Z. Qian, Multi-focus image fusion using an efective discrete wavelet transform based algorithm, Meas. Sci. Rev. 14 (2) (2014) 102–108.
[25] J.J. Lewis, R.J. O'Callaghan, S.G. Nikolov, D.R. Bull, N. Canagarajah, Pixel and region based image fusion with complex wavelets, Inf. Fusion 8 (2) (2007) 119–130.
[26] F. Nencini, A. Garzelli, S. Baronti, Remote sensing image fusion using the curvelet transform, Inf. Fusion 8 (2007) 143–156.
[27] S.F. Ding, X.Y. Zhao, H. Xu, Q.B. Zhu, Y. Xue, NSCT-PCNN image fusion based on image gradient motivation, IET Comput. Vis. 12 (4) (2018) 377–383.
[28] J.J. Wang, Q. Li, Z.H. Jia, N. Kasabov, J. Yang, A novel multi-focus image fusion method using PCNN in nonsubsampled contourlet transform domain, Optik 126 (20) (2013) 2508–2511.
[29] J.M. Xia, Y.M. Chen, A.Y. Chen, Y.C. Chen, Medical image fusion based on sparse representation and PCNN in NSCT domain, Comput. Math. Methods Med. (2018) 1–12.
[30] S.Q. Liu, T. Zhang, H.L. Li, J. Zhao, H.Y. Li, Medical image fusion based on nuclear norm minimization, Int. J. Imaging Syst. Technol. 25 (4) (2015) 310–316.
[31] R. Shen, I. Cheng, A. Basu, Cross-scale coefficient selection for volumetric medical image fusion, IEEE Trans. Biomed. Eng. 60 (4) (2013) 1069–1079.
[32] S. Li, X. Kang, J. Hu, Image fusion with guided filtering, IEEE Trans. Image Process. 22 (7) (2013) 2864–2875.
[33] S. Das, M.K. Kundu, NSCT-based Multimodal medical image fusion using pulse-coupled neural network and modified spatial frequency, Med. Biol. Eng. Comput. 50 (10) (2012) 1105–1114.
[34] S. Liu, J. Zhao, M. Shi, Medical image fusion based on improved sum modified Laplacian, Int. J. Imaging Syst. Technol. 25 (3) (2015) 206–212.
[35] Y. Liu, X. Chen, R.K. Ward, Image fusion with convolutional sparse representation, IEEE Signal Process. Lett. 23 (12) (2016) 1882–1886.
[36] L.D. Li, W.S. Lin, X.S. Wang, G.B. Yang, K. Bahrami, A.C. Kot, No reference image blur assessment based on discrete orthogonal moments, IEEE Trans. Cybern. 46 (1) (2016) 39–50.
[37] L.D. Li, H.C. Zhu, G.B. Yang, J.S. Qian, Referenceless measure of blocking artifacts by tchebichef kernel analysis, IEEE Signal Process. Lett. 21 (1) (2014) 122–125.
[38] L.D. Li, Y. Zhou, W.S. Lin, J.J. Wu, X.F. Zhang, B.J. Chen, No-reference quality assessment of deblocked images, Neurocomputing 177 (2016) 572–584.
[39] R. Hassen, Z. Wang, M.M. Salama, Objective quality assessment for multiexposure multifocus image fusion, IEEE Trans. Image Process. 24 (9) (2015) 2712–2723.
[40] K. Gu, J. Zhou, J.F. Qiao, G.T. Zhai, W.S. Lin, A.C. Bovik, No-Reference quality assessment of screen content pictures, IEEE Trans. Image Process. 26 (8) (2017) 4005–4018.
[41] L.D. Li, Y. Zhou, K. Gu, W.S. Lin, S.Q. Wang, Quality assessment of DIBR-Synthesized images by measuring local geometric distortions and global sharpness, IEEE Trans. Multimed. 20 (4) (2018) 914–926.
[42] Y. Zhou, L.D. Li, S.Q. Wang, J.J. Wu, Y.M. Fang, X.B. Gao, No-Reference quality assessment for view synthesis using gog-based edge statistics and texture naturalness, IEEE Trans. Image Process. 28 (9) (2019) 4566–4579.
[43] K.D. Ma, Z.F. Duanmu, H. Yeganeh, Z. Wang, Multi-exposure image fusion by optimizing a structural similarity index, IEEE Trans. Comput. Imaging 4 (1) (2018) 60–72.
[44] A. Sholehkerdar, J. Tavakoli, Z. Liu, In-depth analysis of Tsallis entropy-based measures for image fusion quality assessment, Opt. Eng. 58 (3) (2019) 033102, 1-16.
[45] X.K. Min, G.T. Zhai, K. Gu, Y.T. Liu, X.K. Yang, Blind image quality estimation via distortion aggravation, IEEE Trans. Broadcast. 64 (2) (2018) 508–517.
[46] K.D. Ma, W.T. Liu, K. Zhang, Z.F. Duanmu, Z. Wang, W.M. Zuo, End-to-End blind image quality assessment using deep neural networks, IEEE Trans. Image Process. 27 (3) (2018) 1202–1213.