Available online at www.sciencedirect.com
ScienceDirect
Image and Vision Computing 27 (2009) 1421–1432
www.elsevier.com/locate/imavis

A new automated quality assessment algorithm for image fusion ☆

Yin Chen *, Rick S. Blum

ECE Department, Lehigh University, Bethlehem, PA 18015-3084, USA

Received 10 February 2007; received in revised form 14 November 2007; accepted 17 December 2007

Abstract

Automated image quality assessment is highly desirable to evaluate the performance of various image fusion algorithms for night vision applications. In this paper we propose a perceptual quality evaluation method for image fusion which is based on human visual system (HVS) models. Our method assesses the image quality of a fused image using the following steps. First, the source and fused images are filtered by a contrast sensitivity function (CSF) after which a local contrast map is computed for each image. Second, a contrast preservation map is generated to describe the relationship between the fused image and each source image. Finally, the preservation maps are weighted by a saliency map to obtain an overall quality map. The mean of the quality map indicates the quality of the fused image. Experimental results compare the predictions made by our algorithm with human perceptual evaluations for several different parameter settings in our algorithm. The most popular existing algorithms are also evaluated. For some specific parameter settings, we find our algorithm provides better predictions, which are more closely matched to human perceptual evaluations, than the existing algorithms. The evaluations focus on the night vision application, but the algorithm we propose is applicable to other applications also.

Keywords: Image fusion; Image quality; Human visual system model; Contrast

1. Introduction

Image fusion provides improved situational awareness of the real world. These images could be obtained from different frames taken by a single camera or sensor, or from different sensors, possibly employing different modalities. Various fusion approaches have been developed in recent years [1–5], and a fundamental issue of image fusion is to evaluate the performance of a fusion scheme.

Traditionally, the quality of a fused image is judged subjectively by a number of human observers. However, this method is time consuming and expensive, and it can not be utilized in a real time manner. A mathematical performance measure able to predict image fusion quality automatically would be highly desirable. Sadjadi [6] investigated a set of approaches to compare the performance of typical image fusion algorithms when certain prior information is available. Some newly developed quality metrics examine the information preserved in the process of using the source images to generate the fused image. For example, Petrovic and Xydeas [7] proposed an objective edge based performance measure which computes the amount of edge information that is transferred from source images to the fused image. Mutual information is employed for assessing fusion quality in a paper by Qu et al. [8]. Piella and Heijmans [9] proposed a new quality metric for image fusion based on research by Wang and Bovik [10] on a structural similarity (SSIM) measure. The measure carries out a quantitative correlation analysis between the source images and the fusion image. However, none of these quality measures take into consideration human visual perception in the metric development.

☆ Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-06-2-0020. The views and conclusions contained in the document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation heron.

* Corresponding author. Tel.: +1 610 758 4447.
E-mail address: yic3@lehigh.edu (Y. Chen).

0262-8856/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.imavis.2007.12.002

The human visual system (HVS) is enormously complex, and there are several models developed to simulate how human beings perceive the quality of an image. Usually a model for the vision process includes contrast calculations, contrast sensitivity filtering, and mutual masking. The human visual response depends more on the contrast of stimuli than on the absolute value of luminance. Furthermore, the sensitivity to stimulus at different spatial frequencies is different which could be modeled by a Contrast Sensitivity Function (CSF). Masking refers to the reduced ability to detect a stimulus on a spatially or temporally complex background.

Recently, Chen and Varshney [11] proposed an image quality metric based on the human visual system. The method employs the CSF on the entire image and then considers the local spatial information transfer on a region-by-region basis. The idea of using models of the human visual system is very interesting and deserved further study. The particular approach given in [11] has a few drawbacks. First the quality value being computed is not bounded, which makes it hard to compare its performance with other measures. Further, we cannot generally say which values indicate good image quality. Secondly, the difference between the source image and the fused one is calculated using the absolute intensity value in a region, while the human eyes are more sensitive to local contrast change.

In this paper, we present a new perceptual quality measure for image fusion, which employs the major features in a human visual system model. The paper is organized as follows. In Section 2, we review the structure and characteristics of our human visual system model. In Section 3, we describe our new quality assessment approach. The experimental results are given in Section 4. Finally conclusions are given in Section 5.

2. The human visual system

2.1. Overview of human visual system

Vision is the most essential of our senses, and human perceptual behavior is a very complicated procedure, which has received a great deal of study. A number of vision models have been proposed to simulate the human perceptual response. We describe the fundamental components of these vision models and extract the important features for use in image quality measure development.

When applying the concepts of human perception to image quality metric development, the following characteristics of the HVS are extremely important [12].

- The human visual response depends more on the contrast of stimuli than on the absolute value of intensity.
- The visual system is not equally sensitive to all stimuli. There are a number of inherent limitations with respect to the visibility of stimuli. Imperfect optics coupled with neural interactions can produce a non-uniform frequency response that is known as a contrast sensitivity function.
- Masking is a critical phenomena in vision, and it refers to the reduced ability to detect a stimulus on a spatially or temporally complex background.

2.2. Fundamentals of human perceptual modeling

A vision model takes advantage of the lower level physiology of the visual system to determine the visual sensitivity [13]. In the following, we consider some parts of the general model in [13] which are useful in building a quality assessment procedure.

2.2.1. Contrast conversion

To illustrate the main idea of contrast, consider an image consisting of a background of uniform intensity I, which has a patch with a different intensity $I + \delta I$ inserted at the center of the image. The observer is asked to determine the point at which they can first detect the patch while increasing δI. For a wide range of background intensity values, the ratio of noticeable difference δI to I is a constant. This relationship, known as Weber's Law, can be expressed as:

$$\frac{\delta I}{I} = K.$$

Weber's Law suggests a commonly used contrast measure which is computed by the ratio of local variations to the surrounding luminance.

2.2.2. Contrast sensitivity function

The contrast sensitivity function describes how sensitive human eyes are to the various frequencies of visual stimuli. Contrast sensitivity is defined as the inverse of the contrast threshold, where the contrast threshold is defined as the minimum contrast necessary for an observer to discriminate a change in intensity. One method to determine the contrast threshold is described in [14]. The contrast of one image is changed and the observer is asked to choose the image with higher contrast. The amount of contrast added to the original stimuli in order to have a 75% detection probability is called the contrast threshold. This procedure is repeated for images with sine waves of different spatial frequencies and the resulting curve is called the spatial contrast sensitivity function. Fig. 1 shows one of the resulting curves which we call an empirical contrast sensitivity function. It turns out that the human eye is sensitive to a limited range of frequencies.

2.2.3. Mutual masking

Masking is a very important phenomena in vision and in image processing which describes interactions between stimuli. Masking occurs when a stimulus that is visible by itself cannot be detected due to the presence of another stimulus. Masking affects are usually quantified

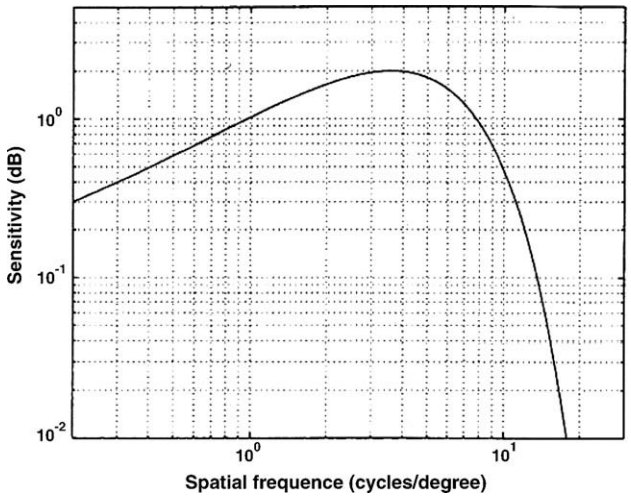


Fig. 1. Spatial contrast sensitivity function.

by measuring the detection threshold for a target stimulus when it is superimposed on another stimulus (the masker) with varying contrast [14].

3. Proposed quality measure for fusion

In this section, we propose a new perceptual image fusion quality assessment method, motivated by the process of human vision modeling. As in many fusion applications, we are not able to obtain the ideal image for comparison reference. Therefore, the information in the input images I_A, I_B which is transferred to the fused image I_F is what we attempt to measure to indicate the fusion quality. Our method includes the following procedure: filtering using a specified contrast sensitivity function; the local contrast computation; the contrast preservation calculation; the saliency map generation; and the global quality measure computation. We describe each step in the following.

3.1. Contrast sensitivity filtering

First, all the images are filtered by an empirical CSF. Let $\mathcal{I}_{\mathcal{A}}(u,v)$ denote the Fourier transform of the source image $I_A(x,y)$, where (x,y) denotes a pixel location. The CSF filtered response $\widetilde{\mathcal{I}}_{\mathcal{A}}(u,v)$ is obtained from

$$\widetilde{\mathcal{I}}_{\mathcal{A}}(u,v) = \mathcal{I}_{\mathcal{A}}(u,v)S(r) \tag{1}$$

where $S(r)$ is the CSF in polar coordinates such that $r = \sqrt{u^2+v^2}$. There are a number of CSFs being proposed for different applications. In our study, we investigate the following choices:

1. Mannos & Sakrison:

$$S_M(r) = 2.6(0.0192 + 0.114r)\mathrm{e}^{-(0.114r)^{1.1}} \tag{2}$$

2. Barton:

$$S_B(r) = r\mathrm{e}^{-0.25r} \tag{3}$$

3. DoG:

$$S_D(r) = \mathrm{e}^{-(r/f_0)^2} - a\mathrm{e}^{-(r/f_1)^2} \tag{4}$$

where f_0 and f_1 are called center and surround cut off spatial frequencies, and a is the amplitude parameter. We use $f_0 = 15.3870$, $f_1 = 1.3456$, and $a = 0.7622$ as suggested by Watson [15].

The inverse Fourier transform is applied to the filtered responses $\widetilde{\mathcal{I}}_{\mathcal{A}}(u,v)$, $\widetilde{\mathcal{I}}_{\mathcal{B}}(u,v)$, and $\widetilde{\mathcal{I}}_{\mathcal{F}}(u,v)$ to yield $\widetilde{I}_A(x,y)$, $\widetilde{I}_B(x,y)$, and $\widetilde{I}_F(x,y)$.

3.2. Local contrast computation

A local contrast calculation is applied to $\widetilde{I}_A$, $\widetilde{I}_B$, and $\widetilde{I}_F$ to give a contrast estimate for each pixel in these images. Local contrast measures have received extensive study, see for example work by Peli, Lubin, Winkler, and other researchers [12,16–18]. In our research we consider what might be called the three most popular contrast calculation methods: a local RMS contrast definition [18], a local band-limited contrast measure [16,17], and an isotropic local contrast measure [12].

3.2.1. Local RMS contrast definition

Consider a pixel with coordinates (x,y) and define a round window of radius p containing N pixels centered at (x,y). To the ith pixel in the window we assign the weight

$$\omega_i = 0.5\left(\cos\left(\frac{\pi}{p}\sqrt{(x_i-x)^2+(y_i-y)^2}\right)+1\right). \tag{5}$$

Define $\overline{I} = \sum_{i=1}^{N}\omega_i I_i$, then the local contrast function of an input image I is defined as [18]:

$$C^{rms}(x,y) = \sqrt{\frac{1}{\sum_{i=1}^{N}\omega_i}\sum_{i=1}^{N}\omega_i\frac{(I_i-\overline{I})^2}{(\overline{I}+I_{dark})^2}} \tag{6}$$

where $C^{rms}(x,y)$ is the RMS (root mean square) contrast at pixel location (x,y) and I_{dark} is the "dark light" parameter chosen to be 1 cd/m^2 based on human photopic intensity discrimination data. Raj [18] pointed out that this parameter has little effect on the measured contrast because the mean luminance of most images are typically larger than 1 cd/m^2.

3.2.2. Local band-limited contrast measure

Peli [16] proposed a measure where the contrast at each point in the image is calculated separately to address the variation of contrast across the image called the *local band-limited contrast*. Define a filter bank of band-pass filters as $\psi_1, \psi_2, \ldots, \psi_M$. Define a corresponding set of low-pass filters $\phi_1, \phi_2, \ldots, \phi_M$. Peli's contrast calculation for an image I can be expressed as:

$$C^P(x,y) = \frac{\psi_j(x,y)\bigstar I(x,y)}{\phi_{j+1}(x,y)\bigstar I(x,y)} \tag{7}$$

where ★ denotes the two dimensional convolution in the (x,y) domain.

Considering the band-pass filters of a pyramid transform, which can be obtained as the difference of two neighboring low-pass filters, Peli's contrast definition can be also written as:

$$C^P(x,y) = \frac{(\phi_j(x,y) - \phi_{j+1}(x,y))\bigstar I(x,y)}{\phi_{j+1}(x,y)\bigstar I(x,y)} = \frac{\phi_j(x,y)\bigstar I(x,y)}{\phi_{j+1}(x,y)\bigstar I(x,y)} - 1. \tag{8}$$

A common choice for ϕ_j would be a Gaussian kernel

$$G_j(x,y) = \frac{1}{(\sqrt{2\pi}\sigma_j)^2} e^{-\frac{x^2+y^2}{2\sigma_j^2}} \tag{9}$$

with a standard deviation equal to $\sigma_j = 2^j$.

Lubin [17] slightly modified this definition and proposed:

$$C^L(x,y) = \frac{(\phi_j(x,y) - \phi_{j+1}(x,y))\bigstar I(x,y)}{\phi_{j+2}(x,y)\bigstar I(x,y)}. \tag{10}$$

The use of $\phi_{j+2}\bigstar I(x,y)$ rather than $\phi_{j+1}\bigstar I(x,y)$ is justified [12].

3.2.3. Isotropic local contrast measure

Winkler introduced an isotropic contrast measure based on analytic filters [12]. His method is defined in the frequency domain in polar coordinates (r,φ). Define a filter bank of band-pass filters with different orientations as $\widehat{\Psi}_{11}, \widehat{\Psi}_{12}, \ldots, \widehat{\Psi}_{KM}$, where the first index denotes the frequency band and the second denotes the different orientations. Define the associated isotropic low-pass filters as $\hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_M$. Then an isotropic contrast measure for an image I at level j is defined as:

$$C^W(x,y) = \frac{\sqrt{2\sum_k |\Psi_{jk}\bigstar I(x,y)|^2}}{\phi_j \bigstar I(x,y)}. \tag{11}$$

3.3. Contrast preservation calculation

Here, we compute the amount of information preserved in the inherent transformation from the input image I_A to the fused image I_F. If we use one of the contrast calculations just described on image $\widetilde{I}_A$, we call the result the contrast map C_A. Similarly the contrast maps obtained from $\widetilde{I}_B$ and $\widetilde{I}_F$ are called C_B and C_F. Let C'_A be the masked contrast map obtained from empirically modeling masking as described in [19]. Then

$$C'_A = \frac{k(C_A)^p}{h(C_A)^q + Z} \tag{12}$$

where k, h, p, q, and Z are real scalar parameters that determine the shape of the non-linearity of the masking function. The information preservation value is then computed by:

$$Q_{AF}(x,y) = \begin{cases} \frac{C'_A(x,y)}{C'_F(x,y)} & \text{if } C'_A(x,y) < C'_F(x,y) \\ \frac{C'_F(x,y)}{C'_A(x,y)} & \text{otherwise.} \end{cases} \tag{13}$$

The value of $Q_{AF}(x,y)$ is in the range of [0,1]. $Q_{AF}(x,y) = 1$ stands for the information preserved from A to F without loss, while $Q_{AF}(x,y) = 0$ corresponds to complete loss of information.

3.4. Saliency map generation

Some researchers use edge strength to represent local saliency [20]. To identify the visual saliency based on a human observer we consider the masked contrast maps. The saliency map λ_A for input image I_A is defined as:

$$\lambda_A(x,y) = \frac{C'^2_A(x,y)}{C'^2_A(x,y) + C'^2_B(x,y)}. \tag{14}$$

Here, C'_A is the masked contrast map obtained from (12). Similarly, the saliency map λ_B is computed for I_B, where A and B denote the two sensor images to be fused.

3.5. Global quality map

The global quality map is computed as

$$Q_C(x,y) = \lambda_A(x,y)Q_{AF}(x,y) + \lambda_B(x,y)Q_{BF}(x,y), \tag{15}$$

where the value of each pixel of Q_C is between 0 and 1.

To obtain a single value indicating the quality of the fused image, we simply take the mean of the quality map over the entire image to obtain

$$Q = \overline{Q_C(x,y)}. \tag{16}$$

According to which CSF was used among S_M, S_B, and S_D, the quality measure being computed is denoted as Q_M, Q_B, and Q_D, respectively.

4. Experimental results

We conducted experimental tests to evaluate the image fusion algorithms and the proposed quality measure. We test the: (1) additive (ADD), (2) discrete wavelet transform (DWT), (3) Filter-Subtract-Decimate pyramid (FSD), (4) Laplacian pyramid (LAP), (5) Morphological pyramid (Morph), and (6) Shift invariant DWT (SiDWT) fusion schemes. The methods (2) to (6) are known as multiresolution decomposition based (MDB) fusion schemes. Table 1 shows the specifications of the tested fusion schemes using the terminology defined in [24]. To assess the performance of our quality metrics Q_M, Q_B, and Q_D, we implemented some existing metrics for the purpose of comparison. There are a number of quality measures discussed in the litera-

Table 1
Fusion schemes (and their specifications) to be tested

Fusion scheme	Reference	MSD level	Activity measure	Grouping method	Combining method	Weights ((%) IITV/IR)	Verification method
ADD	[22]	None	None	None	None	70/30	None
DWT	[23,24]	4	Coefficient based	None	Weighted average	70/30	None
FSD	[26,24]	4	Coefficient based	None	Weighted average	70/30	None
LAP	[25,24]	4	Coefficient based	None	Weighted average	70/30	None
Morph	[3,24]	4	Coefficient based	None	Weighted average	70/30	None
SiDWT	[4,24]	4	Coefficient based	None	Weighted average	70/30	None

ture. Some of them have already been found to be unsuitable for image fusion applications [21], and some of them require prior information on specific targets [6]. Therefore we consider the objective edge based measure [20], the universal index based measure [9], and the information based measure [8] which are developed for image fusion and also utilize features of both the fused image and source images.

4.1. Some existing quality measures

4.1.1. Objective edge based measure (QE)

Xydeas and Petrovic [20] addressed an objective fusion performance measure associated with edge intensity and orientation. The measure is obtained by evaluating the amount of edge information that is transferred from source images to the fused image. A Sobel edge operator is applied to yield edge strength $g(x,y)$ and orientation $\alpha(x,y) \in [0,\pi]$ for each pixel of the image. Then the relative strength and orientation values, $G^{AF}(x,y)$ and $\Phi^{AF}(x,y)$, of input image A with respect to fused image F are defined as

$$G^{AF}(x,y) = \begin{cases} \frac{g_F(x,y)}{g_A(x,y)} & \text{if } g_F(x,y) > g_A(x,y) \\ \frac{g_A(x,y)}{g_F(x,y)} & \text{otherwise} \end{cases} \tag{17}$$

and

$$\Phi^{AF}(x,y) = 1 - \frac{|\alpha_A(x,y) - \alpha_F(x,y)|}{\pi/2}. \tag{18}$$

The edge preservation values Q^{AF} from input image A to fused result F is formed by the product of a sigmoid mapping function of the relative strength and orientation factors. Some constants κ, σ, and Γ determine the shape of the sigmoid mapping as

$$Q^{AF}(x,y) = \frac{\Gamma_g \Gamma_a}{(1 + e^{\kappa_g(G^{AF}(x,y)-\sigma_g)})(1 + e^{\kappa_\alpha(\Phi^{AF}(x,y)-\sigma_\alpha)})}. \tag{19}$$

In our test, $\kappa_g = -15, \sigma_g = 0.5, \Gamma_g = 1.0006$ and $\kappa_\alpha = -22$, $\sigma_\alpha = 0.8, \Gamma_\alpha = 1.0123$ have been employed. The overall objective quality measure QE is obtained by weighting the normalized edge preservation values of both input images as

$$QE = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} Q^{AF}(x,y)\omega^A(x,y) + Q^{BF}(x,y)\omega^B(x,y)}{\sum_{i=1}^{N}\sum_{j=1}^{M}(\omega^A(x,y) + \omega^B(x,y))}. \tag{20}$$

In general the weights $\omega^A(x,y)$ and $\omega^B(x,y)$ are a function of edge strength. The range of QE is between 0 and 1, while 0 indicates the complete loss of source information, and 1 means the best possible fusion performance is obtained.

4.1.2. Universal index based measure (UI)

Piella and Heijmans [9] proposed a new quality metric for image fusion based on research by Wang and Bovik [10] on a structural similarity (SSIM) measure. Given two discrete-time non-negative signals $s = (s_1, \ldots, s_n)$ and $t = (t_1, \ldots, t_n)$, we let μ_s, σ_s^2 and σ_{st} be the mean of s, the variance of s, and the covariance of s and t, respectively. Then the structural similarity index between signal s and t is defined as

$$\text{SSIM} = \frac{\sigma_{st}}{\sigma_s \sigma_t} \cdot \frac{2\mu_s\mu_t}{\mu_s^2 + \mu_t^2} \cdot \frac{2\sigma_s\sigma_t}{\sigma_s^2 + \sigma_t^2}. \tag{21}$$

The new quality index based on the SSIM measure gives an indication of how much of the salient information contained in each of the input images has been transferred into the fused image. First we calculate $\text{SSIM}(a, f \mid w)$ and $\text{SSIM}(b, f \mid w)$ which are the structural similarity measures between the input images and the fused image in a local window w. Then a normalized local weight $\lambda(w)$ is computed from the local saliency of the input images to indicate the relative importance of the source images. The fusion quality index is calculated by

$$UI = \frac{1}{|W|}\sum_{w \in W}(\lambda(w)\text{SSIM}(a,f|w) + (1 - \lambda(w))\text{SSIM}(b,f|w)) \tag{22}$$

where W is the set of all windows and $|W|$ is the cardinality of W.

4.1.3. Information based measure (MI)

Mutual information has been employed as a means of assessing image fusion quality. Define the joint histogram of the fused image F and the source image $A(B)$ as $p_{FA}(f,a)(p_{FB}(f,b))$. Then the mutual information between the fused image and each source image is [8]

$$I_{FA}(f,a) = \sum_{f,a} p_{FA}(f,a)\log_2 \frac{p_{FA}(f,a)}{p_F(f)p_A(a)} \tag{23}$$

and

$$I_{FB}(f,b) = \sum_{f,b} p_{FB}(f,b)\log_2 \frac{p_{FB}(f,b)}{p_F(f)p_B(b)} \quad (24)$$

respectively. Image fusion performance is measured by the value of

$$MI = I_{FA}(f,a) + I_{FB}(f,b). \quad (25)$$

where a larger measure implies better image quality.

4.2. Visual inspection

We mainly focus on night vision image test sets. Figs. 2(a)–(h) and 3(a)–(h) show two examples of sets of source (Figs. 2 and 3(a)–(b)) and fused images (Figs. 2 and 3(c)–(h)). To validate the performance of our new proposed image quality measure, we conducted an informal subjective test among a small group of observers. In this experiment, each observer was shown 28 sets of night vision source images and the corresponding fused results. The positions of the fused images were placed in random sequence in the tests to avoid any positional selection bias. The test environment was kept constant during the tests. The observers viewed the images on a 21 in. ViewSonic LCD monitor. Before taking the test, the observers were briefly instructed in the practical goals of pixel-level image fusion. Then the observers were asked to give an opinion score between 1 and 10, where "1" stands for worst perceptual quality, and "10" implies the best perceptual quality. The observers performed the tests individually with no time limit. The rank scores were recorded anonymously on a form given to each subject.

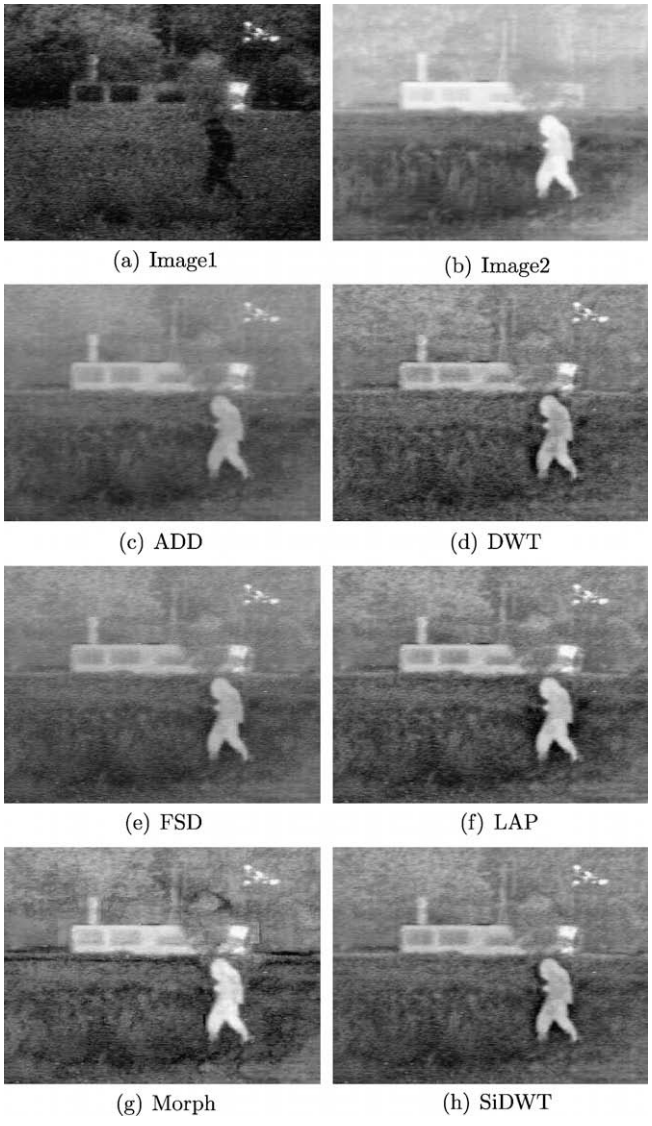


Fig. 2. One example of night vision images and fused results.

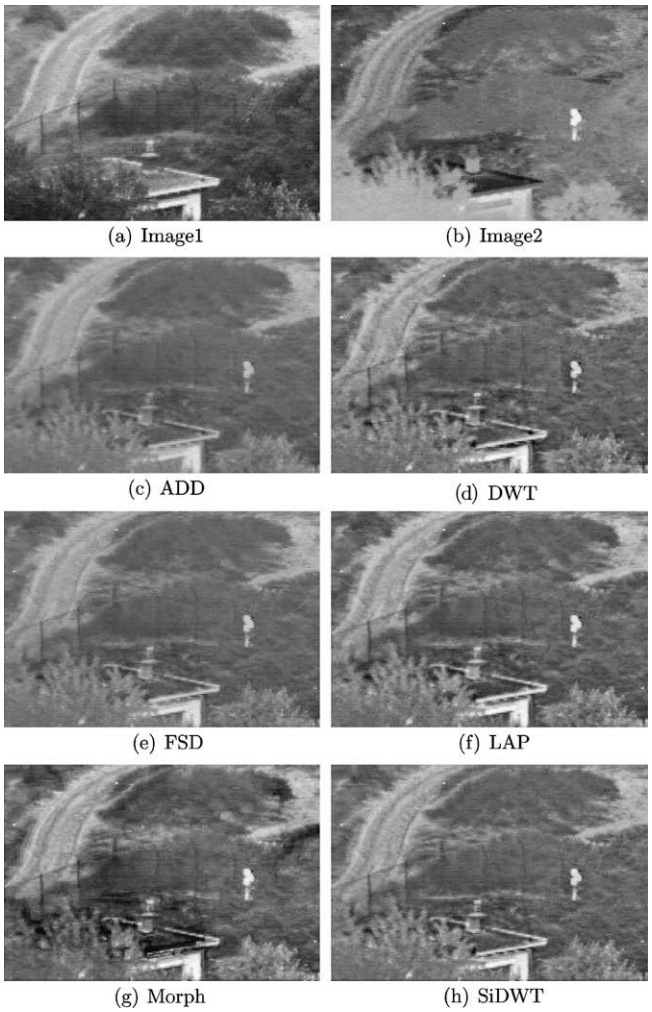


Fig. 3. One example of night vision images and fused results.

Tables 2 and 3 show the mean and standard deviation of the opinion scores for the tested image fusion algorithms. The scores are normalized, divided by 10, so that all scores fall between 0 and 1 which is consistent with the range of the quality measures. As Table 2 shows, generally speaking the observers prefer the MDB fusion methods to ADD fusion. We note that the standard deviations shown in Table 3 are all about 0.1 which means the rankings are fairly consistent. The preferred MDB method seems to vary with the different images based on Table 2. Some methods always rank at the top (LAP, SiDWT) or at the bottom (ADD) but other methods vary from image to

Table 2
Mean of normalized opinion scores over observers for 28 test images

Image	ADD	DWT	FSD	LAP	Morph	SiDWT
1	0.51111	0.71333	0.63111	**0.83667**	0.68556	0.77222
2	0.61333	0.72556	0.58889	**0.81111**	0.71111	0.73333
3	0.56111	0.73111	0.70444	**0.76111**	0.73333	0.69667
4	0.56111	0.57778	0.61889	0.67222	0.65000	**0.69667**
5	0.61222	0.73444	0.67778	**0.79222**	0.58000	0.78333
6	0.68000	0.70556	0.69333	**0.80222**	0.65889	0.76111
7	0.55333	0.74444	0.68556	0.74111	0.72444	**0.77444**
8	0.58667	0.74556	0.58444	0.75889	0.64444	**0.77222**
9	0.64444	0.74444	0.73222	**0.75444**	0.71667	0.73667
10	0.71778	0.75778	0.73444	**0.82222**	0.76111	0.79111
11	0.62778	0.70222	0.70111	**0.75889**	0.64444	0.74222
12	0.57889	0.81778	0.68778	0.82111	0.75111	**0.83889**
13	0.62667	0.74222	0.69444	0.77444	0.67778	**0.79667**
14	0.57778	**0.75556**	0.62000	0.73667	0.63667	0.69444
15	0.59556	0.69222	0.68667	**0.82778**	0.64444	0.74444
16	0.65111	0.79333	0.66778	**0.81333**	0.71111	0.77667
17	0.59000	0.71667	0.69222	**0.79667**	0.67556	0.73111
18	0.58889	**0.83778**	0.68556	0.82889	0.72111	0.71111
19	0.63333	0.75222	0.71111	**0.84556**	0.77667	0.77222
20	0.64000	0.73667	0.65333	**0.78889**	0.74444	0.78778
21	0.66222	0.72444	0.68667	**0.78889**	0.72111	0.74667
22	0.54444	0.64222	0.65000	**0.69222**	0.65889	0.66778
23	0.61778	0.67444	0.66000	**0.79333**	0.66778	0.74778
24	0.59444	0.63556	0.61444	**0.76778**	0.62111	0.70556
25	0.60222	**0.72778**	0.62889	0.71889	0.69667	0.72444
26	0.61000	0.71500	0.66556	0.72667	0.72111	**0.72889**
27	0.60000	0.70222	0.68444	**0.79667**	0.65778	0.74778
28	0.62444	**0.70889**	0.63889	0.70556	0.66222	0.70222

The bolds denote the best fusion method in each row.

Table 3
Standard deviation of normalized opinion scores over observers for 28 test images

Image	ADD	DWT	FSD	LAP	Morph	SiDWT
1	**0.06009**	0.09605	0.09597	0.08646	0.11069	0.09052
2	0.07817	0.09084	0.11057	0.10833	**0.05465**	0.10897
3	0.08580	0.09373	**0.08487**	0.10240	0.12217	0.09987
4	**0.04859**	0.09244	0.07960	0.08899	0.07071	0.06103
5	**0.07345**	0.09812	0.07855	0.10952	0.07906	0.11180
6	0.08170	0.09289	0.09811	**0.08012**	0.12232	0.11816
7	0.07141	0.09167	0.07667	0.10529	0.13510	**0.05812**
8	0.09474	0.07485	**0.07055**	0.12434	0.08819	0.07949
9	**0.07265**	0.09825	0.07396	0.09989	0.13153	0.12884
10	**0.09859**	0.09960	0.11642	0.11211	0.11667	0.11385
11	0.07546	0.12568	0.11297	0.07721	**0.03909**	0.09162
12	0.08069	0.07902	0.07396	**0.04343**	0.12404	0.06009
13	0.09260	0.09705	0.07683	0.08443	0.10220	**0.07053**
14	0.06667	0.10725	0.10943	0.11927	**0.04528**	0.08368
15	**0.09787**	0.10803	0.10828	0.10341	0.13694	0.12551
16	**0.09493**	0.11587	0.11054	0.09862	0.12937	0.11147
17	**0.06633**	0.10308	0.10604	0.09798	0.10667	0.10228
18	**0.07407**	0.09808	0.08946	0.09842	0.09493	0.12484
19	**0.07906**	0.11692	0.08937	0.08413	0.10404	0.13055
20	0.09028	0.10747	0.09987	0.09532	**0.07401**	0.09080
21	0.09338	0.09422	0.12835	0.08580	**0.03951**	0.08646
22	0.07683	0.09135	0.10607	**0.07612**	0.12088	0.11065
23	**0.07049**	0.08876	0.10012	0.08761	0.11065	0.08969
24	**0.06346**	0.11359	0.11293	0.11443	0.13642	0.12105
25	**0.08318**	0.11487	0.11308	0.12283	0.13181	0.09289
26	**0.07681**	0.08116	0.07780	0.11180	0.08667	0.11952
27	**0.07071**	0.10569	0.09735	0.12470	0.10269	0.10256
28	0.09684	0.09662	**0.08937**	0.10737	0.09510	0.09550

The bolds denote the best fusion method in each row.

image in the rankings. It is reasonable that some fusion methods may be better for certain source images, but may not work well for other source images when the scene in the image changes.

4.3. Fusion performance prediction for night vision images

For each test image set, the quality measures Q_M, Q_B, Q_D, QE, UI, and MI are computed. Peli's band-limited contrast measure in (8) is adopted to generate the contrast map. To simplify the computation, we consider only one level of contrast, and the low-pass filter is taken as a Gaussian convolution kernel with the standard deviation $\sigma_1 = 2$. The parameters in (12) are fixed at $k = 1$, $h = 1$, $p = 2.4$, $q = 2$, and $Z = 0.0001$ for the experiment.

To compare the predicted values obtained from the quality measures with the human evaluation scores, we compute the correlation between them. Consider a particular metric. For each test image set i ($i = 1$ to 28), let Q_{ij} denote the quality values predicted for fusion method j. Similarly, let H_{ij} denote the mean of the normalized opinion score (Table 2) for test image i and fusion method j. The correlation coefficient between $(Q_{i1}, \ldots, Q_{i6})$ and $(H_{i1}, \ldots, H_{i6})$ is:

$$r_i = \frac{\sum_{j=1}^{6}\left(Q_{ij} - \frac{\sum_{j=1}^{6} Q_{ij}}{6}\right)\left(H_{ij} - \frac{\sum_{j=1}^{6} H_{ij}}{6}\right)}{\sqrt{\sum_{j=1}^{6}\left(Q_{ij} - \frac{\sum_{j=1}^{6} Q_{ij}}{6}\right)^2\left(H_{ij} - \frac{\sum_{j=1}^{6} H_{ij}}{6}\right)^2}}. \quad (26)$$

To get further insight we consider statistics across the image sets. To understand how the correlation varies over the different image sets, it is useful to be able to count the percentage of image sets for which the correlation coefficient takes on certain values. We can do this easily by using the notion of cumulative probability distribution (cdf). Thus we define the cdf as the percentage of cases for which we obtain a correlation coefficient less than some value.

Fig. 4 compares the cdfs for different quality measures. It shows that Q_D correlates best with the human evaluation for the test images since its cdf falls to the right (closest to 1) of all others. This indicates the correlation values for this method are largest. Q_D and QE outperform the other measures with much higher correlations to human evaluations. Furthermore among the 28 test images, Q_D obtains a higher average correlation than QE (0.8946 compared to 0.7993), and a smaller standard deviation (0.0644 compared to 0.0959) when such statistics are computed across the image test sets.[1] Q_M, Q_B, and UI show good correlations only for a few test images. The percentage of good correlations is quite low and the standard deviation of the correlation is large when statistics are computed across the image test sets. For example, from Fig. 4 we notice that about 40%

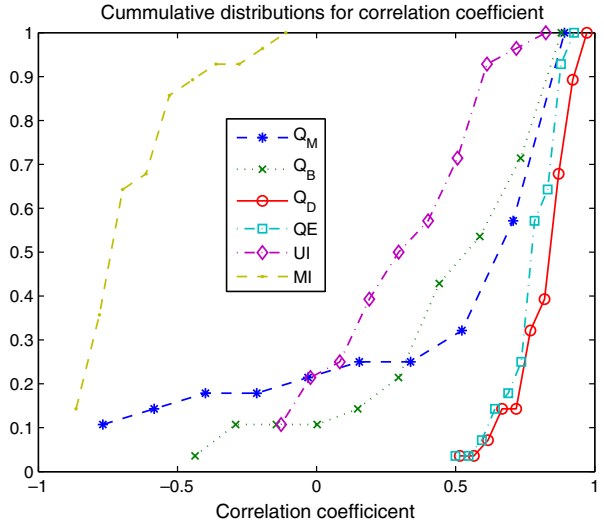


Fig. 4. Comparison of the cumulative distribution functions of the correlation coefficient for different quality measures.

of Q_M predictions are well correlated to the opinion scores if we consider a good correlation to be 0.75. About 10% of Q_M predictions show no correlation with the opinion scores ($r_i < 0$). The MI measure shows no apparent correlation to the human evaluations for the test images with most of the correlation coefficients less than zero.

We also measured the root mean square error (RMSE) between the normalized mean opinion score and the quality measure predictions for each test image set. This error would tell the accuracy of the quality measure predictions when compared to the normalized average human rank scores. The RMSE for test image set i is defined as:

$$\mathrm{RMSE}_i = \sqrt{\frac{1}{6}\sum_{j=1}^{6}(Q_{ij} - H_{ij})^2}. \quad (27)$$

Fig. 5 shows the cumulative probability distribution of RMSE for different quality measures. From Fig. 5, it appears that the proposed methods Q_M, Q_B, and Q_D yield smaller RMSE than the existing methods QE, UI, and MI. Q_D yields a smaller mean RMSE than Q_M and Q_B (0.0637 compared to 0.0703 and 0.0704), but a larger standard deviation (0.0394 compared to 0.0247 and 0.0214). This larger standard deviation is seen in Fig. 5. From Fig. 5 we see that about 20% of Q_D predictions give slightly larger RMSE than Q_M and Q_B predictions.

Next, we examine the preference ranking among the tested fusion algorithms for each test image set. We compare the ranking difference between human evaluations and those obtained by different quality measures. Let the image fusion algorithms be arranged in a fixed sequence (1) ADD; (2) DWT; (3) FSD; (4) LAP; (5) Morph; (6) SiDWT. First we sort the normalized opinion score vector $\vec{H}_i = \{H_{i1}, H_{i2}, \ldots, H_{i6}\}$ for image set i in ascending order, and return an array giving the rank order which we call

[1] The cdf does give an indication of the standard deviation from how spread out along the x-axis the curves are.

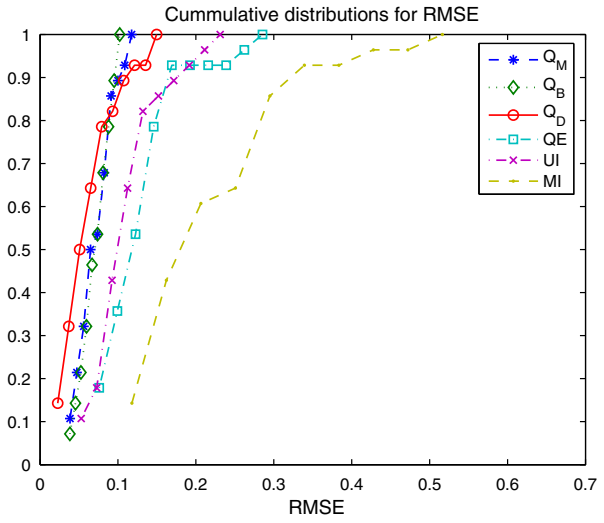


Fig. 5. Comparison of the cumulative distribution functions of RMSE for different quality measures.

$\vec{O}_{H_i} = \{O_{H_i1}, \ldots, O_{H_i6}\}$. Thus, O_{H_i3} gives the rank order for the 3rd element of $\vec{H}_i$. If $O_{H_i3} = 4$ then the third element of $\vec{H}_i$ is the fourth largest element. Thus if $H_{i6} < H_{i4} < H_{i5} < H_{i3} < H_{i1} < H_{i2}$, then $\vec{O}_{H_i} = \{5, 6, 4, 2, 3, 1\}$. Consider a particular quality measure. Next we sort the quality values for image set i and obtain a rank order array $\vec{O}_{Q_i} = \{O_{Q_i1}, \ldots, O_{Q_i6}\}$. Ideally, if the predictions from the quality measure perfectly match the human evaluation scores, $\vec{O}_{Q_i}$ would be same as $\vec{O}_{H_i}$. In other words, the corresponding elements in $\vec{O}_{Q_i}$ and $\vec{O}_{H_i}$ are identical. We record the number of places where an element in $\vec{O}_{Q_i}$ mismatches the respective element in $\vec{O}_{H_i}$ and denote it as D_i.

$$D_i = \sum_{j=1}^{6} \mathbb{I}(O_{H_ij} \neq O_{Q_ij}), \tag{28}$$

where $\mathbb{I}$ is an indicator function with $\mathbb{I}(\text{true}) = 1$ and $\mathbb{I}(\text{false}) = 0$. Table 4 shows the average of D_i for different quality measures.

Although Q_D leads to the best performance in Table 4, it is not close to zero (since zero means the order between human evaluations and quality measure predictions perfectly match). We note that if some algorithm is ranked incorrectly, then so is some other algorithms. This may explain why the rank difference for Q_D is closer to 2 than 1. Further thought will convince one that some misrankings will effect more than two algorithm ranks. For example the proper rankings could be {1,2,3,4,5,6}. However the quality measure could have the same order with only one algorithm rank shifted. In this case we get the ranking {6,1,2,3,4,5}. This would lead to all 6 rankings being mismatched.

Table 4
Comparison of the average number of rank order differences between human evaluations and quality measure predictions

Q_M	Q_B	Q_D	QE	UI	MI
4.1071	4.2500	2.7143	3.6071	5.1786	5.5000

Table 5
Comparison of the average number of misclassifications between human evaluations and quality measure prediction

Q_M	Q_B	Q_D	QE	UI	MI
1.2500	1.6786	0.5714	0.6071	2.1071	2.5357

Since some of the fused methods yield similar fused images, it may be more reasonable to categorize the fused methods into groups instead of giving them strict preference orders. For each test image set, we find the minimum in the normalized opinion score vector $\vec{H}_i = \{H_{i1}, H_{i2}, \ldots, H_{i6}\}$ and denote it as M_{H_i}. Let τ_H be a positive value. We categorize the fused methods into two groups: "good" and "not good". If $H_{ij} \geqslant M_{H_i} + \tau_H$, the corresponding fusion method j is classified as a "good" fusion method. If $H_{ij} < M_{H_i} + \tau_H$, the corresponding fusion method j is classified as a "not good" fusion method. Consider a particular quality measure. We then classify the fusion methods based on the quality values $\vec{Q}_i = \{Q_{i1}, Q_{i2}, \ldots, Q_{i6}\}$. Let M_{Q_i} denote the minimum in $\vec{Q}_i$. Let τ_Q be a positive value. The fusion method j is considered to be "good" when $Q_{ij} \geqslant M_{Q_i} + \tau_Q$, and to be "not good" when $Q_{ij} < M_{Q_i} + \tau_Q$. When the classification for the quality measure is different from the human evaluation, we consider the case a misclassification. Table 5 lists the average number of misclassifications over 28 test images. Again Q_D yields the best performance with QE close behind. Further, for Q_D and QE less than one fused algorithm of 28 tested, will be misclassified.

4.4. Fusion performance evaluation with additive Gaussian noise

In this experiment, we add Gaussian noise with the variance σ_n^2 to the source images. Since the additive fusion method computes the weighted average of two source images, it can reduce the effect of the introduced noise at the cost of some loss of high frequency information. On the other hand, some non-weighted averaging methods are developed for keeping the high frequency information from the source images. In noisy environments, these non-weighted averaging methods may not perform well since the noise could be preserved or even enlarged in the fused result. As a particular example of such a non-weighted averaging method, we consider the method in [23,24] which we call the DWT fusion method.

Due to the lack of a "true scene" image, an undistorted ideal image which shows all objects of interest in all source images with perfect clarity, in night vision applications, we use the image set in Fig. 6 in this experiment. Fig. 6(a) shows a known test image. From it, we create two out-of-focus images by blurring the original image with a Gaussian smoothing kernel. Fig. 6(b) and (c) show the

(a) Original image (b) focus on the left (c) focus on the right

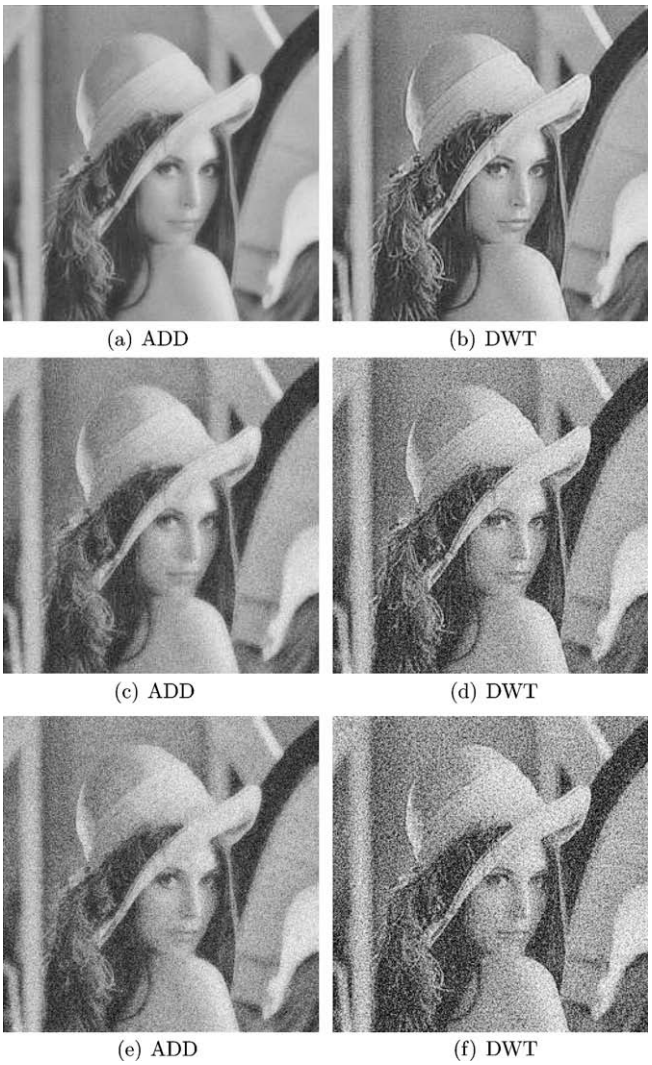
Fig. 6. Test image set.

source images with blurring in the right half and left half of the images respectively. We add some Gaussian noise into the source images and apply the fusion algorithms to the noisy source images. Fig. 7(a)–(f) shows some examples of the fused images obtained from the noisy source images. From our tests, when there is no Gaussian noise or a small amount of noise, DWT method yields better fused results than weighted averaging (ADD) fusion. When the level of noise increases, the ADD fusion rule provides better fused results because the noise is smoothed or averaged out a bit.

Generally, the quality measure should be decreasing when we add more noise into the source images and we do see this in the results. Also, as justified in Fig. 7(a) and Fig. 7(b), when the amount of noise is relatively small, for example the variance is less than 0.0005 in this case, the DWT fusion leads to better fusion performance than additive fusion. When the variance of noise is larger (around 0.0005 in this case), the additive fusion (Fig. 7(c) and Fig. 7(e)) outperform over DWT fusion (Fig. 7(d) and Fig. 7(f)). Fig. 8(a), (b), and (c) shows the performance predictions reported by the quality measures Q_M, Q_B, and Q_D. It seems that all three do predict a crossover point between ADD and DWT fusion performance when Gaussian noise is added into the images. Using the original image as a reference image, or "true scene" image, in this case, we calculate the peak signal-to-noise ratio (PSNR) [11] between the fused image and the reference in Fig. 9(a) for both ADD and DWT fusion. The crossover point appears at a noise variance around 0.0005. One might suggest that Q_M and Q_B lead to better predictions since they obtain the closest crossover point to that suggested in the PSNR plot in Fig. 9(a).

We implemented the edge-based quality measure QE for the noisy image fusion, and Fig. 9(b) shows the result. We see that QE fails to predict the preference of additive fusion when the noise increases. QE indicates that DWT fusion is better than ADD fusion no matter how much noise is added to the source images, which is inconsistent with the human observations.

We conducted the same test using other images, and the experiments show that our proposed method usually better matches human judgements on the image quality when compared to the QE metric.



(a) ADD (b) DWT

(c) ADD (d) DWT

(e) ADD (f) DWT

Fig. 7. Some fusion results for multi-focus images with Gaussian noise: Left column images are fused by taking weighted average [22], right column images are fused by DWT [23,24]. From top to bottom, the variance of the noise are: $\sigma_n^2 = 0.0005$(top), $\sigma_n^2 = 0.005$(middle), $\sigma_n^2 = 0.01$(bottom).

4.5. Discussion

In this section, we have examined the proposed quality measures in several different ways. We compare the

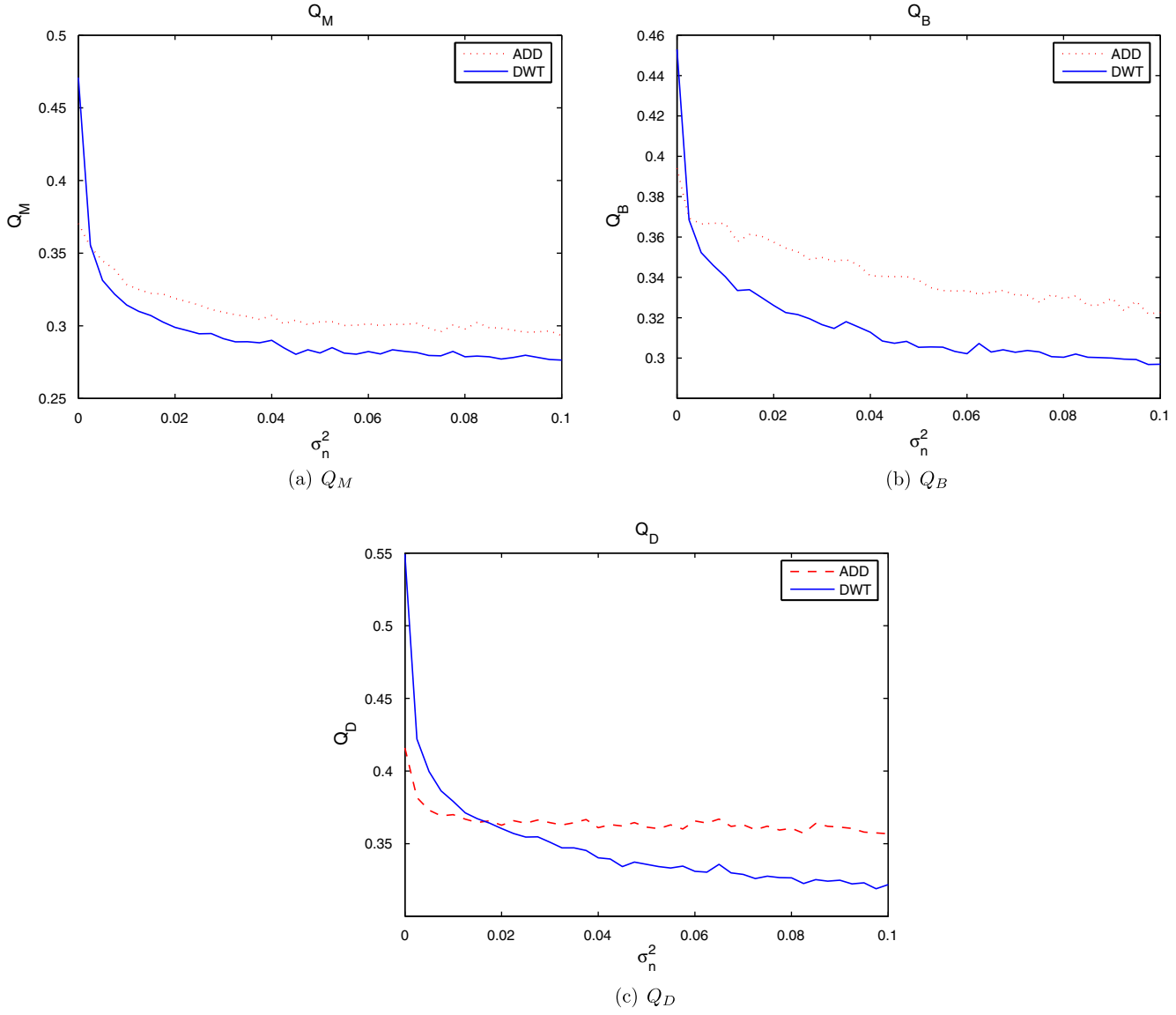


Fig. 8. The performance of proposed quality metric for noisy source images.

correlation coefficient, RMSE, preference orders and good–bad classification between the quality measure predictions and human evaluations. Note that the correlation coefficient in (26) is invariant to any scalar multiplication of $\{Q_{i1},\ldots,Q_{i6}\}$ or $\{H_{i1},\ldots,H_{i6}\}$. Thus a large correlation coefficient indicates the shape of $\{Q_{i1},\ldots,Q_{i6}\}$ and $\{H_{i1},\ldots,H_{i6}\}$ match well. Thus this criterion considers the relative ranking across different fusion methods. RMSE, on the other hand, shows how the absolute magnitudes of the quality measure differ from the human evaluation scores. This criterion would help to judge the accuracy of the numerical value of each quality prediction. Based on these criteria, Q_D generally performs the best for the night vision images we tested. We also examined how well the algorithms are ranked and classified by the different quality measures. For these evaluations we also find Q_D performs better than the other quality measures.

We utilized Peli's band-limited contrast measure in (8) to develop the quality measures, and there are other contrast calculation methods discussed in Section 3. We also tested them in our experiments and the results show that Peli's band-limited contrast measure leads to overall better performance than other contrast measures. We also conducted some experiments for noisy source images in multi-focus fusion applications, and in that case our proposed quality measures matches the human evaluation better than QE metric.

5. Summary

In this paper, we consider the human vision system and propose a new quality measure based on the human visual system modeling. We compared the proposed measures Q_M, Q_B, and Q_D with some existing quality measures QE,

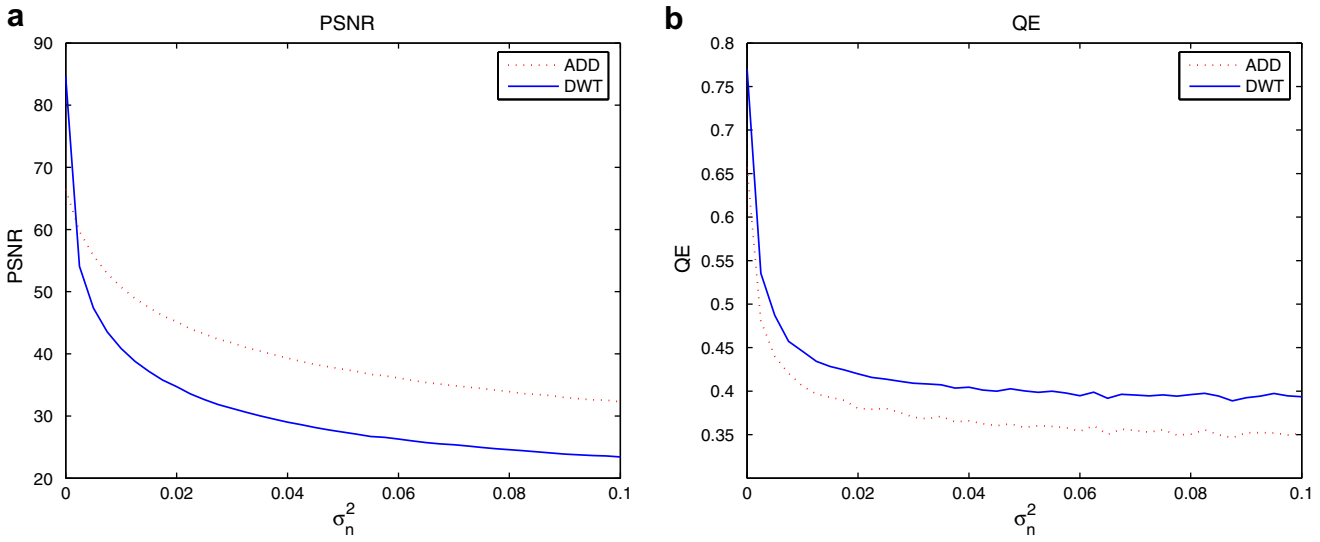


Fig. 9. The performance of PSNR (a) and QE (b) for the noisy source images.

UI, and *MI*. The experiments show that Q_D outperforms the others for night vision test images.

References

[1] S. Frechette, V.K. Ingle, Gradient based multifocus video image fusion, in: IEEE Conference on Advanced Video and Signal Based Surveillance, 2005, pp. 486–492.

[2] P.K. Varshney, H. Chen, L.C. Ramac et al, Registration and fusion of infrared and millimeter wave images for concealed weapon detection, in: Proceedings of International Conference on Image Processing, 1999, pp. 532–536.

[3] A. Toet, A morphological pyramidal image decomposition, Pattern Recognition Letters 9 (1989) 255–261.

[4] O. Rockinger, Image sequence fusion using a shift invariant wavelet transform, in: Proceedings of the International Conference on Image Processing, 1997, pp. 288–291.

[5] C. Pohl, J.L. Van Genderen, Multisensor image fusion in remote sensing: concepts, methods and applications, Interantional Journal of Remote Sensing 19 (5) (1998) 823–854.

[6] F. Sadjadi, Comparative image fusion analysis, in: Joint IEEE International Workshop on OTCBVS, 2005, pp. 8–15.

[7] V. Petrovic, C. Xydeas, On the effects of sensor noise in pixel-level image fusion performance, in: Proceedings of the 3rd International Conference on Information Fusion, 2000, pp. 14–19.

[8] Guihong Qu, Dali Zhang, Pingfan Yan, Information measure for performance of image fusion, Electronics Letters 38 (7) (2002) 313–315.

[9] G. Piella, A new quality metric for image fusion, in: Proceedings of International Conference on Image Processing, 2003, pp. 173–176.

[10] Zhou Wang, Alan C. Bovik, A universal image quality index, IEEE Signal Processing Letters 9 (3) (2002) 81–84.

[11] H. Chen, P.K. Varshney, A perceptual quality metric for image fusion based on regional information, Proceedings of SPIE 5831 (2005) 34–45.

[12] S. Winkler, Digital Video Quality Vision Models and Metrics, John Wiley & Sons, 2005.

[13] T. Pappas, R. Safranek, Perceptual criteria for image quality evaluation, in: A. Bovik (Ed.), Handbook of Image and Video Processing, Academic Press, 2000.

[14] G.E. Legge, J.M. Foley, Contrast masking in human vision, Journal of Optical Society of America 70 (1980) 1458–1470.

[15] A.B. Watson, A.J. Ahumada Jr., A standard model for foveal detection of spatial contrast, Journal of Vision 5 (9) (2005) 717–740.

[16] E. Peli, Contrast in complex images, Journal of the Optical Society of America A 7 (10) (1990) 2032–2040.

[17] J. Lubin, A visual discrimination model for imaging system design and evaluation, in: E. Peli (Ed.), Vision Models for Target Detection and Recognition, World Scientific Publishing, Singapore, 1995, chapter 10.

[18] R.G. Raj, W.S. Geisler, R.A. Frazor, A.C. Bovik, Natural contrast statistics and the selection of visual fixations, International Conference on Image Processing 3 (September) (2005) 1152–1155.

[19] J.M. Foley, Human luminance pattern-vision mechanisms: masking experiments require a new model, Journal of Optical Society of America 11 (6) (1994) 1710–1719.

[20] C.S. Xydeas, V. Petrovic, Objective image fusion performance measure, Electronics Letters 36 (4) (2000) 308–309.

[21] Yin Chen, R.S. Blum, Experimental tests of image fusion for night vision, in: Proceeding of the 8th International Conference on Information Fusion, 2005.

[22] E.J. Bender, C.E. Reese, G.S. van der Wal, Comparison of additive image fusion versus feature-level image fusion techniques for enhanced night driving, in: Proceedings of SPIE, Vacuum and Solid State Photoelectronic Imagers, Detectors, and Systems, vol. 4796, 2002, pp. 140–151.

[23] T. Huntsberger, B. Jawerth, Wavelet based sensor fusion, in: Proceedings of SPIE, vol. 2059, 1993, pp. 488–498.

[24] Z. Zhang, R.S. Blum, A categorization of multiscale-decomposition-based image fusion schemes with a performance study for a digital camera application, Proceedings of IEEE 87 (8) (1999) 1315–1326.

[25] Peter J. Burt, Edward H. Adelson, The laplacian pyramid as a compact image code, IEEE Transaction on Communications 31 (4) (1983) 532–540.

[26] C.H. Anderson. A filter-subtract-decimate hierarchical pyramid signal analyzing and synthesizing technique. United States Patent 4,718,104, Washington, D.C., 1987.