Information Fusion 14 (2013) 127–135

Contents lists available at SciVerse ScienceDirect

Information Fusion

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

# A new image fusion performance metric based on visual information fidelity

Yu Han [a,b], Yunze Cai [a,*], Yin Cao [a], Xiaoming Xu [a,c]

[a] Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, PR China
[b] Jiangsu Automation Research Institute, Lianyungang 222006, PR China
[c] University of Shanghai for Science and Technology, Shanghai Academy of Systems Science, Shanghai 200093, PR China

ARTICLE INFO

*Article history:*
Received 6 January 2011
Received in revised form 1 August 2011
Accepted 16 August 2011
Available online 25 August 2011

*Keywords:*
Image fusion assessment
Fused image quality
Visual information fidelity
Visual information fidelity for fusion

ABSTRACT

Because subjective evaluation is not adequate for assessing work in an automatic system, using an objective image fusion performance metric is a common approach to evaluate the quality of different fusion schemes. In this paper, a multi-resolution image fusion metric using visual information fidelity (*VIF*) is presented to assess fusion performance objectively. This method has four stages: (1) Source and fused images are filtered and divided into blocks. (2) Visual information is evaluated with and without distortion information in each block. (3) The visual information fidelity for fusion (*VIFF*) of each sub-band is calculated. (4) The overall quality measure is determined by weighting the *VIFF* of each sub-band. In our experiment, the proposed fusion assessment method is compared with several existing fusion metrics using the subjective test dataset provided by Petrovic. We found that *VIFF* performs better in terms of both human perception matching and computational complexity.



## 1. Introduction

Multiple imaging systems are commonly used to improve situational awareness of a complex scene because they capture the scene without losing information and then integrate effective visual information from source images together to produce a high-quality result. This integration process is called fusion, which is accomplished by fusion algorithms. Because an increasing number of fusion approaches have been proposed [1] in recent years, evaluating fusion algorithm performance has become an important issue.

The conventional way to evaluate fusion performance is to score the fused images subjectively by a number of trained observers, which requires significant economic consumption and cannot be used in an automatic system. As a result, objective fusion metric has become the primary focus because it can automatically predict the performance of fusion algorithms.

Mutual information (*MI*) given by Qu et al. [2] was the most commonly used objective metric for image fusion. Hossny et al. [3] revised the *MI* metric and proposed the Normalized *MI* (*NMI*) metric. Both metrics took advantage of information theory and demonstrated the significance of the theory. Xydeas and Petrovic [4] proposed an edge-based fusion performance measure $Q_E$ by calculating edge strength and orientation between the source and fused image. Xydeas's work showed high performance for fusion assessment. The structural similarity index measure (*SSIM*) [8] which showed better performance in image quality assessment was employed for fusion performance metrics. For instance, Piella [5] used weighted *SSIMs* between the source images and fused image in each block to determine the weighted fusion quality (*WFQ*) and edge dependent fusion quality index (*EDFQI*). Cvejic et al. [6] improved Piella's work and determined weighted parameters based on block similarity between two source-fused image pairs. Unlike considering both gray and edge information in Piella's work, Yang's et al. [7] metric only focused on the gray information in images. In Yang's work, *WFQ* was not used to predict fusion performance exclusively, while the maximum *SSIM* of each block was utilized as a substitute for *WFQ* values under certain conditions. Chen and Varshney [9] proposed a new fusion metric based on the human visual system (*HVS*). The method employed the contrast sensitivity function (*CSF*) on the entire image and then used the local spatial information transfer on a region-by-region basis. Chen and Varshney studied the best parameter settings for their algorithm and assessed their algorithm using different types of fused images. Chen and Blum [10] presented a new fusion metric based on *HVS*. Their work did not rely on edge information but rather on local contrast in an empirical *CSF* filtered image. They evaluated their metric using a night vision image test set and achieved high performance.

Because the assessment of an image fusion scheme is strongly correlated to the image quality, the development of image quality has a great impact on fusion metrics. This paper presents a new image fusion metric based on visual information fidelity (*VIF*) that has shown high performance for image quality prediction. Both

* Corresponding author.
*E-mail addresses:* hansymail@gmail.com (Y. Han), yzcai.sjtu@gmail.com (Y. Cai), carrie.caoyin@gmail.com (Y. Cao), xmxu@usst.edu.cn (X. Xu).

1566-2535/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.inffus.2011.08.002

predictive performance and computational complexity are improved in this metric. Section 2 briefly introduces *VIF*, and Section 3 describes the principle and framework of our fusion metric algorithm. Experimental results are discussed in Section 4, while the performance of the proposed algorithm is compared with some commonly used fusion metrics in Petrovic's subjective test database [11]. Finally, conclusions are drawn in Section 5.

## 2. Principle of visual information fidelity

This section reviews visual information fidelity, which is an effective full reference image quality metric based on natural scene statistics (*NSS*) theory. The principle of *VIF* is shown in Fig. 1.

As depicted in Fig. 1, *VIF* first decomposes the natural image into several sub-bands and parses each sub-band into blocks. Then, *VIF* measures the visual information by computing mutual information in the different models in each block and each sub-band. Finally, the image quality value is measured by integrating visual information for all the blocks and all the sub-bands. Here, *VIF* introduces three models to measure the visual information: the Gaussian scale mixture (*GSM*) model, the distortion model and the *HVS* model.

The GSM model [14,15] is a *NSS* model in the wavelet domain. A *GSM* is a random field (*RF*) that can be expressed as a product of two independent *RFs*: Gaussian and scale [13], which is expressed as

$$C_i = s_i U_i \tag{1}$$

In (1), $C_i$ denotes the $i$th *RF* of the reference signal in a sub-band; $s_i$ is the $i$th random positive scalar; $U_i$ is the $i$th Gaussian vector *RF*, and its variance is $C_U$. A *GSM* model is a special pixel model, and *RF* is considered a pixel set in a local block at a sub-band. Thus, we alternate *RF* with blocks to help understand the working principle of the metric in the following.

A distortion model is used to describe the extent to which distortion operators can disturb an image. *VIF* adopt a signal attenuation and additive noise distorted model for each *RF*.

$$D_i = g_i C_i + V_i \tag{2}$$

where $C_i$ denotes the $i$th *RF* of the reference signal, which has same meaning as (1), and $D_i$ denotes the corresponding *RF* in the sub-band in the test image. $g_i$ represents the scalar value, which is determined by distortion Eq. (2), and $V_i$ is a stationary additive zero-mean Gaussian noise field with variance $C_{V_i} = \sigma_{V_i}^2$.

The *HVS* model in *VIF* quantifies the impact of the signal that flows through *HVS*. In *VIF*, *HVS* is modeled to be an additive component in the distortion channel. Sheikh and Bovik [13] marks this *HVS* distortion as visual noise and models it as a stationary white additive zero-mean Gaussian noise in the wavelet domain. Thus, in stationary *RFs*, *HVS* noise can be modeled as noise $N$ and $N'$, which are zero-mean uncorrelated multivariate Gaussians with the same dimensionality as $C_i$

$$E_i = C_i + N$$
$$F_i = D_i + N' = g_i C_i + V_i + N'$$

where $E_i$ and $F_i$ denote the cognitive output of the reference and test images extracted from the brain, respectively; they are created by transferring $C_i$ and $D_i$ through the *HVS* model in one sub-band. *VIF* assumes the covariance of $N$ and $N'$ is the same: $\sigma_N^2 = \sigma_{N'}^2$.

To achieve analytical and computational simplicity, *VIF* assumes $N$, $N'$, $U_i$, $s_i$ and $V_i$ are mutually independent for all the blocks and sub-bands. *VIF* utilizes mutual information $I(C_i,E_i)$ to measure the information that can be extracted from the output of *HVS* when the reference image is being viewed.

$$I(C_i, E_i) = h(C_i + N) - h(N) = \frac{1}{2}\log_2\left(\frac{\left|s_i^2 C_U + \sigma_N^2 I\right|}{\left|\sigma_N^2 I\right|}\right) \tag{3}$$

In addition, information $I(C_i,F_i)$ is measured in the same way when the test image is being viewed.

$$I(C_i, F_i) = h\left(g_i \cdot C_i + V_i + N'\right) - h\left(V_i + N'\right) = \frac{1}{2}\log_2\left(\frac{\left|g_i^2 s_i^2 C_U + \left(\sigma_{V_i}^2 + \sigma_N^2\right) I\right|}{\left|\left(\sigma_{V_i}^2 + \sigma_N^2\right) I\right|}\right) \tag{4}$$

where $|\cdot|$ denotes the matrix determinant, and $I$ is the identity matrix with the same dimensions as $C_U$.

The above discussion only considers the $i$th *RF*, regardless of sub-band. When sub-band information is taken into account, *VIF*, which is defined as two information ratios, can be written as

$$VIF = \frac{\sum_{k\in subband}\sum_b I(C_b, F_b)}{\sum_{k\in subband}\sum_b I(C_b, E_b)} = \frac{\sum_k\sum_b \log_2\left(1 + \frac{g_{k,b}^2 s_{k,b}^2 C_U}{\left(\sigma_{V_{k,b}}^2 + \sigma_N^2\right) I}\right)}{\sum_k\sum_b \log_2\left(1 + \frac{s_{k,b}^2 C_U}{\sigma_N^2 I}\right)} \tag{5}$$

where $k$ and $b$ stand for sub-band and block (*RF*) index, respectively; $g_{k,b}$ is the scalar gain field in the $b$th block at the $k$th sub-band, and $s_{k,b}$ and $C_U$ are defined correspondingly. It is evident that $g_{k,b}$ and $s_{k,b}$ are generalized definitions of $g_i$ and $s_i$ when considering multiple sub-bands.

This implementation of *VIF* is not entirely consistent with the theory. An estimated model replaces the theoretical model in practice. For instance, $s_{k,b}^2 C_U$ is estimated from the local variance of pixels based on maximum likelihood (*ML*) criteria through Eq. (6)

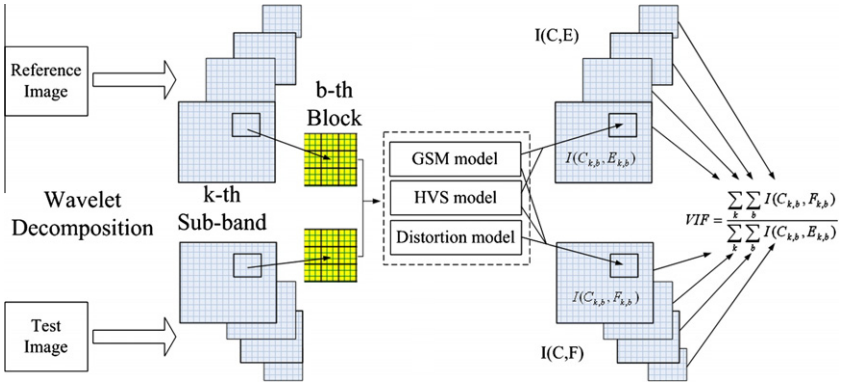


**Fig. 1.** A schematic of *VIF*.

$$s_{k,b}^2 C_U = \left(\sigma_{k,b}^r\right)^2 \tag{6}$$

where $\sigma_{k,b}^r$ stands for the standard deviation of the source image $I_r$ in the $b$th block at the $k$th sub-band, and $g_{k,b}$ in Eq. (2) can be estimated by Eq. (7):

$$g_{k,b} = \sigma_{k,b}^{r,d} \Big/ \left(\sigma_{k,b}^r\right)^2 \tag{7}$$

Thus, $\sigma_{Vk,b}^2$ is estimated by Eq. (8)

$$\sigma_{V_{k,b}}^2 = \left(\sigma_{k,b}^d\right)^2 - g_{k,b}^2 \cdot \left(\sigma_{k,b}^r\right)^2 \tag{8}$$

where $\sigma_{k,b}^{r,d}$ stands for the covariance of source image $I_r$ and test image $I_d$ in the $b$th block at the $k$th sub-band. Actually, *VIF* is treated as a function *VIF*($I_r$, $I_d$) in practice: the inputs of the function are the reference image ($I_r$) and test image ($I_d$), and the output is the assessment value. Therefore, *VIF* for source image $I_r$ and test image $I_d$ is computed according to Eq. (9):

$$VIF(I_r, I_d) = \frac{\sum_k \sum_b \log_2 \left(1 + \frac{g_{k,b}^2 \cdot \left(\sigma_{k,b}^r\right)^2}{\left(\sigma_{k,b}^d\right)^2 - g_{k,b}^2 \cdot \left(\sigma_{k,b}^r\right)^2 + \sigma_N^2}\right)}{\sum_k \sum_b \log_2 \left(1 + \frac{\left(\sigma_{k,b}^r\right)^2}{\sigma_N^2}\right)} \tag{9}$$

## 3. Visual information fidelity for fusion (VIFF) performance metric

In *VIF*, $I(C_i,E_i)$ is equivalent to the Signal Noise Ratio (*SNR*) at the $i$th spatial position of a sub-band when only considering visual noise, while $I(C_i,F_i)$ is another *SNR* at the same position and sub-band with both distortion and visual noise. Thus, the principle of *VIF* can be recounted in the following four steps: first, *VIF* decomposes reference and test images into several sub-bands; second, *VIF* measures spatially local *SNR* with and without distortion information of images at multiple scales; third, *VIF* adds them together to get all the information, which reflects the entire impact on *HVS* with and without considering distortion; fourth, *VIF* assesses the ratio of information with and without distortion as image quality.

In a word, distortion information in spatially local *SNR* is directly related to the image quality in a *VIF* model. Distortion information in spatially local *SNR* reflects how much 'visual information' is in a test image compared with the reference image, whereas 'visual information' here is defined as a good visual response for *HVS*. Based on this idea, a fusion metric is constructed, and its schematic is exhibited in Fig. 2. The main point of the fusion metric is to measure how much 'effective visual information' (*EVI*) in the fused image is extracted from source images, while 'effective visual information' is defined as the maximum visual information of all the source-fused image pairs. In this metric, the *VIF* model is used to extract visual information from source-fused image pairs. Then, *EVI* is determined by selecting visual information from the *VIF* model with the distortion information as a guide. In our work, *EVI* in the local spatial and spectral domains is defined as fusion visual information. Finally, a metric value is calculated by compounding all the fusion visual information into a fusion metric.

Because there are many images in fusion assessment problem, we assume $I_1, I_2, \ldots, I_n$ are n source images, and $I_F$ is the fused image. Visual information without distortion information (*VIND*) for source image $I_i$ and fused image $I_F$ in the $b$th block and $k$th sub-band is defined by Eq. (10)

$$VIND_{k,b}(I_i, I_F) = \frac{1}{2}\log_2 \left(\frac{\left|s_{k,b}^2 C_U + \sigma_N^2 I\right|}{\left|\sigma_N^2 I\right|}\right) \tag{10}$$

Visual information with distortion information (*VID*) for source image $I_i$ and fused image $I_F$ in the same block and sub-band is defined by Eq. (11)

$$VID_{k,b}(I_i, I_F) = \frac{1}{2}\log_2 \left(\frac{\left|g_{k,b}^2 s_{k,b}^2 C_U + \left(\sigma_{V_{k,b}}^2 + \sigma_N^2\right) I\right|}{\left|\left(\sigma_{V_{k,b}}^2 + \sigma_N^2\right) I\right|}\right) \tag{11}$$

In (10) and (11), indexes $b$ and $k$ represent the $b$th block and $k$th sub-band of images, and $g_{k,b}, s_{k,b}, \sigma_{V_{k,b}}^2$ and $\sigma_n^2$ have the same definitions as in *VIF*.

The definitions of *VIND* and *VID* are basically the same as $I(C_i,E_i)$ and $I(C_i,F_i)$ in *VIF* with a slight difference for the fusion metric; *VIF* needs a reference image for computation, while there is no reference image for the fusion problem. Because the reference image is unknown for the fusion problem, the source image is substituted for the reference image as the input, while the fused image is treated as the test image. This substitution is based on the hypothesis that $I(C_i,E_i)$ (*VIND*) and $I(C_i,F_i)$ (*VID*) used in *VIF* reflect the visual information of the test image in comparison with the reference

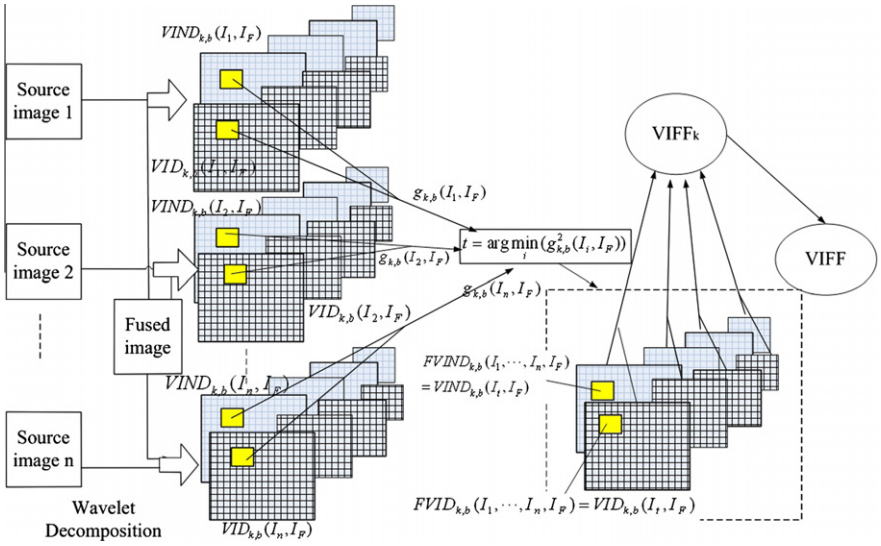


**Fig. 2.** A schematic of a *VIF*-based fusion metric (*VIFF*).

image. Thus, *VIND* and *VID* represent the visual information of a fused image that is extracted from the source image.

Because there are many source images in the fusion problem, the source of *EVI* in the fused image should be determined first. Here, a scalar parameter, $g_{k,b}$, is selected as the source index of *EVI* for the fusion problem. We assume that the *EVI* of the fused image in the $b$th block at the $k$th sub-band is extracted from the $t$th image. $t$ is calculated according to Eq. (12):

$$t = \arg\min_i \left(g^2_{k,b}(I_1, I_F), \ldots, g^2_{k,b}(I_i, I_F), \ldots, g^2_{k,b}(I_n, I_F)\right) \tag{12}$$

In (12), $g^2_{k,b}(I_i, I_F)$ represents the square of the scalar gain $g_{k,b}$ in $VID_{k,b}$ with $I_i$ and $I_F$ as input images. *min* is the information selection principle of the fused image. A small $g^2_{k,b}(I_i, I_F)$ implies that a large quantity of *EVI* is extracted from the $i$th image in the $b$th block at the $k$th sub-band compared with other source images. With the source information index marked, the *EVI* measures, Fusion visual information with distortion (*FVID*) and Fusion visual information without distortion (*FVIND*), in the $b$th block at the $k$th sub-band are defined as follows:

$$\begin{aligned} &FVID_{k,b}(I_1, \ldots, I_n, I_F) = VID_{k,b}(I_t, I_F) \\ &for \quad t = \arg\min_i \left(g^2_{k,b}(I_1, I_F), \ldots, g^2_{k,b}(I_i, I_F), \ldots, g^2_{k,b}(I_n, I_F)\right) \end{aligned} \tag{13}$$

$$\begin{aligned} &FVIND_{k,b}(I_1, \ldots, I_n, I_F) = VIND_{k,b}(I_t, I_F) \\ &for \quad t = \arg\min_i \left(g^2_{k,b}(I_1, I_F), \ldots, g^2_{k,b}(I_i, I_F), \ldots, g^2_{k,b}(I_n, I_F)\right) \end{aligned} \tag{14}$$

Eqs. (13) and (14) reflect the *EVI* selected by the fusion algorithm. For the *EVI* measures, $FVID_{k,b}$ and $FVIND_{k,b}$, are denoted in the local spatial and spectral domains. *VIF* for fusion assessment in the $k$th sub-band is presented as $VIFF_k$; see Eq. (15).

$$VIFF_k(I_1, \ldots I_n, I_F) = \frac{\sum_b FVID_{k,b}(I_1, \ldots, I_n, I_F)}{\sum_b FVIND_{k,b}(I_1, \ldots, I_n, I_F)} \tag{15}$$

Global *VIF* for fusion assessment of *VIFF* is computed by weighting the *VIFF* in each sub-band:

$$VIFF(I_1, \ldots, I_n, I_F) = \sum_k p_k \cdot VIFF_k(I_1, \ldots, I_n, I_F) \tag{16}$$

where $p_k$ is a weighting coefficient. According to *VIF* theory, a high *VIF* yields a high quality test image. Therefore, as *VIFF* increases, the quality of the fused image improves.

Because *VIFF* is based on *VIF*, an estimate of the *VIF* model is adopted to implement *VIFF*. In Section 2, $s^2_{k,b}K_U, g_{k,b}$ and $\sigma^2_{Vk,b}$ are estimated by *ML* criteria in practice. Here, these parameters are estimated using the same method in *VIFF*. $VID_{k,b}(I_i, I_F)$ and $VIND_{k,b}(I_i, I_F)$ are computed as follows:

$$VID_{k,b}(I_i, I_F) = \log_2\left(1 + \frac{g^2_{k,b}(I_i, I_F) \cdot \left(\sigma^i_{k,b}\right)^2}{\left(\left(\sigma^F_{k,b}\right)^2 - g^2_{k,b}(I_i, I_F) \cdot \left(\sigma^i_{k,b}\right)^2 + \sigma^2_N\right)}\right) \tag{17}$$

$$VIND_{k,b}(I_i, I_F) = \log_2\left(1 + \frac{\left(\sigma^i_{k,b}\right)^2}{\sigma^2_N}\right) \tag{18}$$

where $g_{k,b}(I_i, I_F)$ is estimated by

$$g_{k,b}(I_F, I_i) = \sigma^{F,i}_{k,b} / \left(\sigma^i_{k,b}\right)^2 \tag{19}$$

In the above equation, $\sigma^{F,i}_{k,b}$ is the covariance of the source image $I_i$ and fused image $I_F$ in the $b$th block at the $k$th sub-band, $\sigma^i_{k,b}$ and $\sigma^F_{k,b}$ are the standard deviation of source image $I_i$ and fused image $I_F$, respectively, in the $b$th block at the $k$th sub-band.

In summary, the *VIFF* algorithm to evaluate the fusion quality with two source images is given as follows, where $I_1$, $I_2$ and $I_F$ represent two source images and the fused image, respectively. The pixel value of image $I_1$ at row $i$ column $j$ is denoted $I_1\{i,j\}$, and the notation is the same for the other image and matrixes.

1. Initialize $\sigma^2_N$, $VIFF = 0$, 4 value array $p\{\cdot\}$;
2. For $k$ = 1 to 4
   $FVID = 0$;
   $FVIND = 0$;
   Check the size of $I_1$, $I_2$ and $I_F$, assuming the size is $M \times N$
   Construct 2D normal filter $h$ by $k$ with a size of $P \times K$;
3. For $i$ = 1 to $M$
   For $j$ = 1 to $N$
   $\sigma_{1,1} = 0$; $\sigma_{2,2} = 0$; $\sigma_{1,F} = 0$; $\sigma_{2,F} = 0$; $\sigma_{F,F} = 0$;
   $m_1 = 0$; $m_2 = 0$; $m_F = 0$;
4. For $a = -P/2$ to $P/2$
   For $b = -K/2$ to $K/2$
   $\sigma_{1,1} = \sigma_{1,1} + I_1\{i-a,j-b\} \cdot I_1\{i-a,j-b\} \cdot h\{i-a,j-b\}$;
   $\sigma_{2,2} = \sigma_{2,2} + I_2\{i-a,j-b\} \cdot I_2\{i-a,j-b\} \cdot h\{i-a,j-b\}$;
   $\sigma_{F,F} = \sigma_{F,F} + I_F\{i-a,j-b\} \cdot I_F\{i-a,j-b\} \cdot h\{i-a,j-b\}$;
   $\sigma_{1,F} = \sigma_{1,F} + I_1\{i-a,j-b\} \cdot I_F\{i-a,j-b\} \cdot h\{i-a,j-b\}$;
   $\sigma_{2,F} = \sigma_{2,F} + I_2\{i-a,j-b\} \cdot I_F\{i-a,j-b\} \cdot h\{i-a,j-b\}$;
   $m_1 = m_1 + I_1\{i-a,j-b\} \cdot h\{i-a,j-b\}$;
   $m_2 = m_2 + I_2\{i-a,j-b\} \cdot h\{i-a,j-b\}$;
   $m_F = m_F + I_F\{i-a,j-b\} \cdot h\{i-a,j-b\}$;
   End for ($b$); End for ($a$)
   $\sigma_{1,1} = \sigma_{1,1} - m_1 \cdot m_1$;
   $\sigma_{2,2} = \sigma_{2,2} - m_2 \cdot m_2$;
   $\sigma_{1,F} = \sigma_{1,F} - m_1 \cdot m_F$;
   $\sigma_{2,F} = \sigma_{2,F} - m_2 \cdot m_F$;
   $\sigma_{F,F} = \sigma_{F,F} - m_F \cdot m_F$;
5. $g_1 = \sigma_{1,F}/\sigma_{1,1}$;
   $g_2 = \sigma_{2,F}/\sigma_{2,2}$;
   If $|g_1| < |g_2|$
   $VID\{i,j\} = log_2(1 + (g_1 \cdot g_1 \cdot \sigma_{1,1}/$
   $(\sigma_{F,F} - \sigma_{1,F} \cdot \sigma_{1,F}/\sigma_{1,1} + \sigma^2_n)))$
   $VIND\{i,j\} = log_2(1 + \sigma_{1,1}/\sigma^2_n)$
   Else
   $VID\{i,j\} = log_2(1 + (g_2 \cdot g_2 \cdot \sigma_{2,2}/$
   $(\sigma_{F,F} - \sigma_{2,F} \cdot \sigma_{2,F}/\sigma_{2,2} + \sigma^2_n)))$
   $VIND\{i,j\} = log_2(1 + \sigma_{2,2}/\sigma^2_n)$
   End if
6. $FVID = FVID + VID\{i,j\}$;
   $FVIND = FVIND + VIND\{i,j\}$;
   End for ($j$); End for ($i$);
7. Image $I_1$, $I_2$ and $I_F$ are downsampled by 2
   $VIFF = VIFF + p\{k\} \cdot FVID/FVIND$;
   End for ($k$)
8. The whole fusion assessment value is *VIFF*.

In [13,16], filter $h$ is set to a Gaussian filter, and its size and variance decrease as $k$ (sub-band index) increases. Here, the same $h$ is adopted in *VIFF*. The value of $p$ and $\sigma^2_n$ are given in the next section.

Note that steps 3 and 4 are the main part of *VIF* because it computes *VIF* using convolution in the source code [16]. *VIFF* is also implemented through convolution in specific code.

## 4. Validation experiment and discussion

There are several problems in the current comparisons of existing metric algorithms for image fusion. For example, some works [2,3,5–7] compared their evaluation algorithms with readers'

subjective judgments or other common objective image quality assessment methods (such as *PSNR*). The limitations are listed as follows:

1. Direct subjective perception.
   It is not precise to assess fusion metrics using subjective perception. The accuracy and consistency of subjective perception are influenced by many factors, such as illumination and the observer's mood and background. Therefore, subjective perception may vary significantly for the same image. Usually, subjective evaluation of an image is based on the method proposed in ITU Recommendation[1] [20]. Only a value obtained after rigorous subjective evaluation can be used as a subjective rating for an image.
2. Using a reference metric.
   It is not reliable to assess a fusion metric using other image quality metrics as references. According to image quality assessment research [16,18], some commonly used image quality metrics (such as *PSNR*, *SNR*, *Entropy*) do not have good predictive performance for subject perception. There are no existing image quality metrics that are suitable for references because they are not accurate enough.
3. Using one or several pieces of images.
   It is not appropriate to assess fusion metric with only one or several pieces of images. According to the quality metric comparison idea proposed by *VQEG* [17,18], the performance of an image quality assessment algorithm is a statistical standard, and therefore, a proper sample size is necessary. Similarly, in fusion performance assessment, we also need an appropriate number of fusion samples.

Because subjective perception has a great impact on fusion metric, in our experiment, a quality comparison based on *VQEG* is applied. To ensure a large sample size, our metric is tested on a subjective test database issued by Petrovic [11]. The database is composed of 151 monochrome registered image pairs captured by different sensors in real or realistic conditions. For each pair of source images, two fused images are created by different fusion schemes. All 120 pairs of fused images in the database were evaluated by the ITU Recommendation scheme. Subjective evaluation is measured by voting 'preferred image 1', 'preferred image 2' and 'equal preference'. The details of test procedure are described elsewhere [11].

According to Petrovic, for each fused image pair, one subjective preference vector $S$ for each fused pair is shown:

$$S = \begin{cases} [1,0,0] & \textit{prefered image1} \\ [0,1,0] & \textit{prefered image2} \\ [0,0,1] & \textit{equal} \end{cases} \tag{20}$$

where 'preferred image 1' means that image 1 receives more votes. Similarly, objective preference vector $O$, which depends on the objective metrics of each pair of fused images, is shown:

$$O = \begin{cases} [1,0,0]^T & Q_1 > Q_2 \\ [0,1,0]^T & Q_2 > Q_1 \\ [0,0,1]^T & Q_1 \approx Q_2 \end{cases} \tag{21}$$

where $Q_1$ and $Q_2$ are metric scores for the two fused images (schemes 1 and 2). Operator > indicates that metric scores suggest image1 has better quality. Because an explicitly equal score $Q_1 = Q_2$ is almost impossible due to the continuous range of metrics, approximate equality $Q_1 \approx Q_2$ is defined as the case when the absolute error between $Q_1$ and $Q_2$ is less than 0.001.

In Petrovic's work, two performance indicators, subjective relevance (*SR*) and correct ranking (*CR*), were used to compare the fusion metric objectively.

*SR* is used to predict the correlation between subjective and objective results. In our work, *SR* is rewritten as Eq. (22):

$$SR = \frac{\sum_i T_i O_i - \sum_i E S_i^T}{\sum_i T_i S_i^T - \sum_i E S_i^T} \tag{22}$$

where $O_i$ is an option vector of objective metrics for the fused image group $i$, and $S_i$ is the subjective preference vector for the fused image group $i$. $T_i$ is the normalized subject votes vector for the fused image group $i$ whose element sum is one. For example, if the subject vote is [4,10,1], the normalized subject votes vector $T$ is [4/15,10/15,1/15]. $E$ is a special subjective preference vector that has equal preference for each option (the worst case): $E = [1/3, 1/3, 1/3]$. The range of SR is from 0 to 1.

*CR* reflects the prediction accuracy and monotonicity of the fusion metric. It first counts the number of image pairs whose subjective and objective ranking are the same and then computes its ratio for all the pairs:

$$CR = \frac{1}{N} \sum_i S_i O_i \tag{23}$$

where $N$ is the total number of fused image pairs.

As *SR* and *CR* increase, the predictive performance of the image fusion metric improves.

### 4.1. Parameter selection

To ascertain $p_k$ in (10), we optimized *CR* of *VIFF* using the certain optimization method in Petrovic's database. After 100 trials, we found that $p$ must be $[0.465, 0, 0.070, 0.465]$ to get a high *CR* for four sub-bands of *VIFF*.

According to some research [12,13], visual noise $\sigma_n^2$ usually is selected as 0.1 (image scale 0–1). Because *VIFF* and *VIF* are not identical, we measured the influence of $\sigma_n^2$ in the *VIFF*. In our experiment, $\sigma_n^2$ changed from 0.001 to 0.5 with steps of 0.001. *CR* was used as the preference index of $\sigma_n^2$. The main result is shown in Fig. 3.

In Fig. 3, the $x$-axis is $\sigma_n^2$ which varies from 0.001 to 0.5 (image scale 0–1), and the y-axis is *CR*. Based on the results shown in Fig. 3, *VIFF* has a higher predictive performance when $\sigma_n^2$ is in the range $[0.004, 0.006]$. The impact of visual noise for *VIFF* is different from that of *VIF* because of the *EVI* index principle used in *VIFF*. Because *VID* and *VIND* are constructed by information difference (or contrast), the principle that the lower $g_{k,b}$ is chosen as the *EVI* index causes *VID* and *VIND* to decrease. In other words, the visual information contrast decreases. To improve the contrast between *VID* and *VIND*, low visual noise is adopted in *VIFF*. In our work, $\sigma_n^2$ is set to 0.005.

### 4.2. Main results and analysis

We computed the *SR* and *CR* of all fusion assessment metrics introduced in this paper using Petrovic's database. The results are shown in Table 1. In our experiment, all the *SSIM*-based metrics [5–7] did not perform as well as the classic metric $Q^{AB/F}$proposed by Xydeas and Petrovic [4]. Blum's metric showed better predictive ability than $Q^{AB/F}$ in *CR*, while its *SR* was lower than $Q^{AB/F}$. In our experiment, *VIFF* correctly predicted 95 of 120 pairs. It is obvious that *VIFF* has the largest *SR* and *CR* of all other fusion measures using Petrovic's database, which implies its superiority for image fusion assessment.

[1] The ITU Recommendation is widely accepted as an objective and fair methodology to measure the subjective perception of an image.

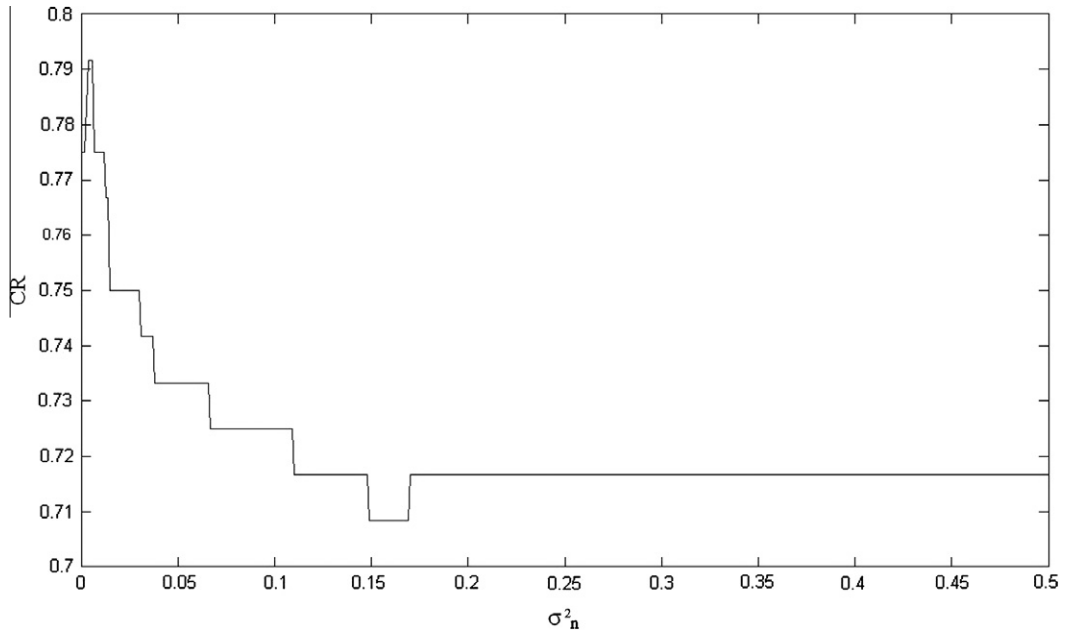


**Fig. 3.** The predictive performance of *VIFF* with different $\sigma_n^2$.

**Table 1**
SR and CR of different fusion metrics.

| | Qu et al. [2] | Hossny et al. [3] | Xydeas and Petrovic [4] | Piella [5] | Cvejic et al. [6] | Yang et al. [7] | Chen and Varshney [9] | Chen and Blum [10] | Our method [21] |
|---|---|---|---|---|---|---|---|---|---|
| *SR* | 0.563 | 0.563 | 0.713 | 0.649 | 0.573 | 0.700 | 0.446 | 0.708 | 0.745 |
| *CR* | 0.650 | 0.650 | 0.725 | 0.717 | 0.616 | 0. 717 | 0.575 | 0.750 | 0.792 |

However, there were still 25 errors for fused images. We believe these errors are caused by some source images in Petrovic's database that do not satisfy the condition of a natural scene statistic. *VIFF* relies on the *GSM* model, which is a method used to model natural scene statistics [19]. According to natural scene statistic theory, the main feature of a natural scene image is that its coefficients distribution is a *power-law-like* distribution.

A sample of natural scene images in Petrovic's database is shown in Fig. 4. Fig. 4 shows a pair of aerial images that contain natural scene. In Fig. 4, images ($a$) and ($b$) are two different source images that must be fused. Curves ($a'$) and ($b'$) are coefficient distributions of images ($a$) and ($b$), respectively. In curve ($a'$), the $x$-axis and $y$-axis stand for coefficient value and its occurrence probability, respectively; the same is true for curve ($b'$). The coefficient distribution is clearly *power-law-like*.

For some hyper-spectral images in Petrovic's database, such as the images in Fig. 5, the coefficient distribution of the image does not satisfy the *power-law-like* distribution condition. In Fig. 5, images ($a$) and ($b$) are two different source images that must be fused. Curves ($a'$) and ($b'$) are the coefficient distributions of images ($a$) and ($b$), respectively. In curve ($a'$), the $x$-axis and $y$-axis stand for the coefficient value and its probability of occurrence, respectively; the same is true for curve ($b'$). The coefficient distribution is not *power-law-like*. As a result, the model assuming natural image statistics becomes unreliable. Thus, the *VIFF* might fail to give a proper assessment. *VIFF* correctly performs fusion assessment in Fig. 4 and fails in Fig. 5.

According to our experience, *VIFF* exhibits good prediction performance in natural scene images and might not be effective for unnatural images. Thus, future research should focus on modeling these unnatural images.

### 4.3. Time complexity discussion and comparison

Predictive ability is not the only criterion we considered. Computational complexity is a concern for real-time processing. As physical memory and other hardware become cheaper than ever before, the space complexity problem seems to be directly resolved by hardware improvement. Thus, only time complexity is discussed in our work.

Because digital images are modeled by a matrix, fusion performance algorithms are often implemented by a combination of matrix operations. The time required for a matrix operation depends on the number of iterations, while the computational burden of an algorithm is determined by the part that contains the most iterations or recurrences. Common operations (e.g., addition of the Hadamard product) have a lower computational burden than high complexity operations (e.g., iterated-based operations, such as convolution or the Fourier transform). Therefore, high complexity operations generate a high computational burden for an algorithm.

Based on the above analysis, we measure the computational burden of fusion metrics using the following three steps:

1. Decomposing the fusion metric into many parts based on their principles;
   According to our research, many metric algorithms could be decomposed as a combination of *SSIM* computation, edge computation, *CSF* filter computation and contrast computation. Some algorithms limit their computation in uncovered local blocks, while most algorithms work through covered local block computation (or convolution).

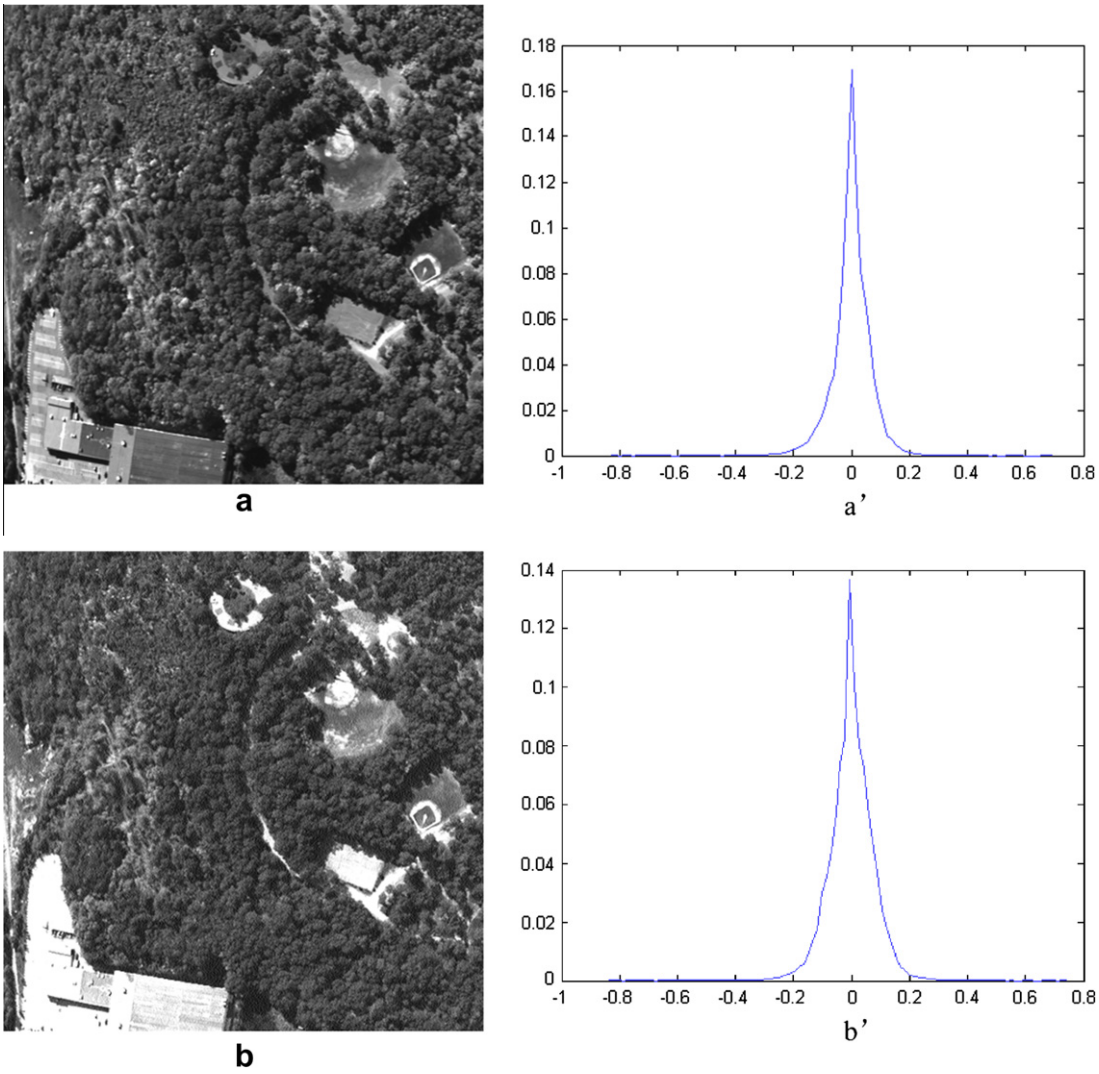


**Fig. 4.** natural scene images ($a$ and $b$) and their coefficient distributions ($a'$ and $b'$) (sample 23 in Petrovic's database). In $a'$ and $b'$, the $x$-axis stands for coefficients, and the $y$-axis shows the probability of occurrence.

2. If uncovered local block computation (*UCLBC*) exists, convert it into covered local block computation (*CLBC*) with the same time consumption;
   Local block computation (*LBC*) is a common technique to extract image local information based on the block character. If every pixel of an image is the block center, it is defined as a covered local block computation (*CLBC*) because the blocks centered by adjacent pixels intersect with each other. However, uncovered local block computation (*UCLBC*) only involve non-intersection block results. Thus, the time consumption of *UCLBC* is lower than *CLBC* because fewer pixels are processed. Fig. 6 describes the difference between covered local block and uncovered local block computation when the block size is $3 \times 3$.
   Note that *CLBC* could be programmed as either a single or a combination of convolution operations and its time consumption is proportional to the image size. *UCLBC* could be transmitted to *CLBC* first then downsampled, while its effective computation is equal to *CLBC* on a downsampled image. Thus, the time consumption of *UCLBC* could be the same as *CLBC* on the downsampled image; the downsampled image is determined by the detailed information in the blocks and the computation.
3. Counting the minimum number of high complexity operators needed in each *CLBC* part of the metric.

As mentioned above, *SSIM* computation, edge computation, contrast computation and *CSF* filter computation are common parts in many fusion metrics. According to one paper [8], one *SSIM* computation requires at least 5 convolution operations. The edge operation requires 2 convolution operations (horizontal and vertical Sobel operators), while contrast computation could also be implemented by convolution combinations based on their definition. We believe that time consumption between *CSF* filter computation, which works in spectrum space and requires the Fourier transform, and convolution operation could be same. Otherwise, computation can be changed because of the consistency between spectrum filtering and spatial convolution.

Finally, as the number of high complexity operators (or convolutions) used in the metric decreases, the metric executes more quickly. The following are examples of metric analysis for the two-source image fusion problem:

For Piella's metric [5]

(1) Generate edge map for all the input and fused images. Because the edge map can be computed by convoluting the image with horizontal and vertical Sobel operators, there are 6 convolution operators (2 convolutions $\times$ 3 images) used in edge computation.

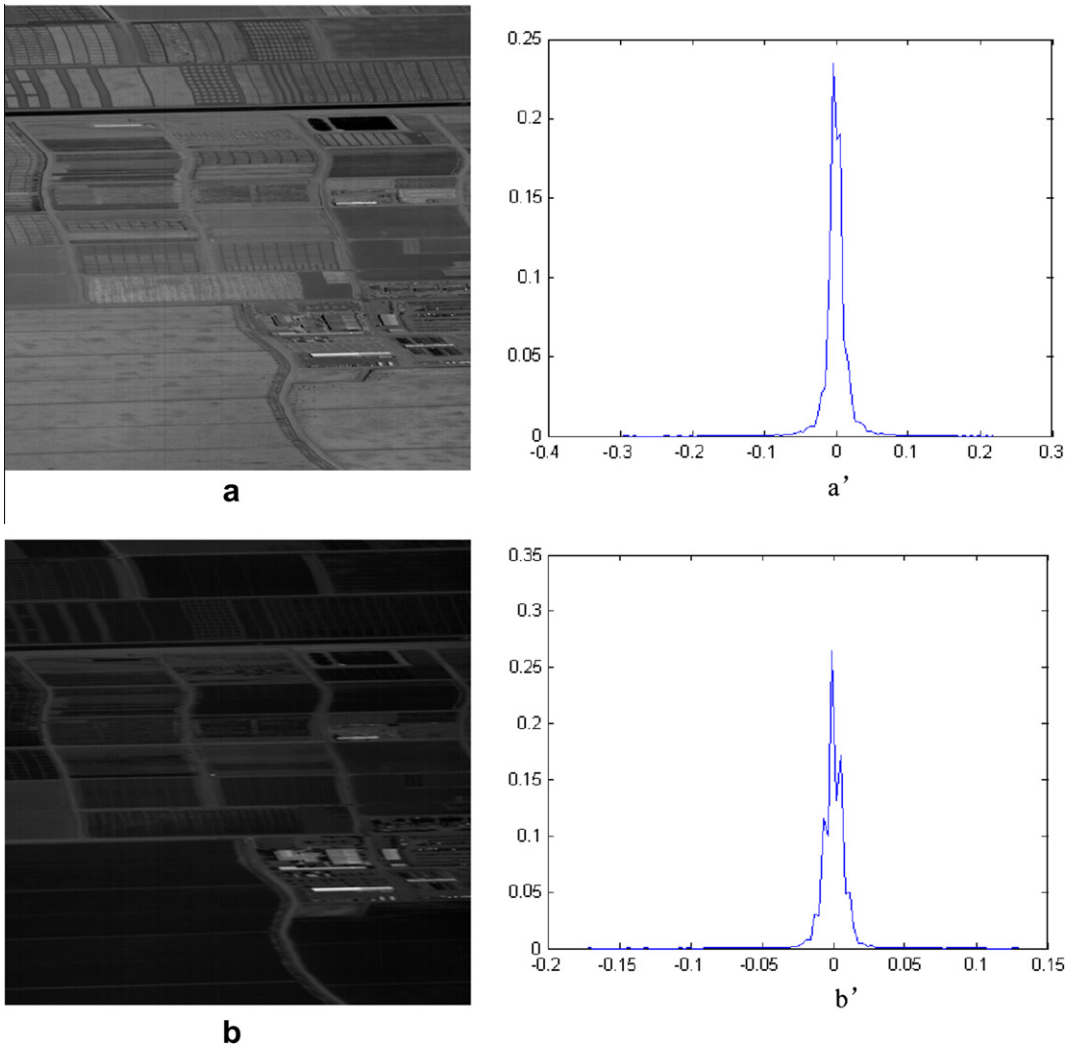


**Fig. 5.** unnatural images ($a$ and $b$) and their coefficients distribution ($a'$ and $b'$) (sample 50 in Petrovic's database). In $a'$ and $b'$, the $x$-axis stands for coefficients, and the $y$-axis shows the probability of occurrence.

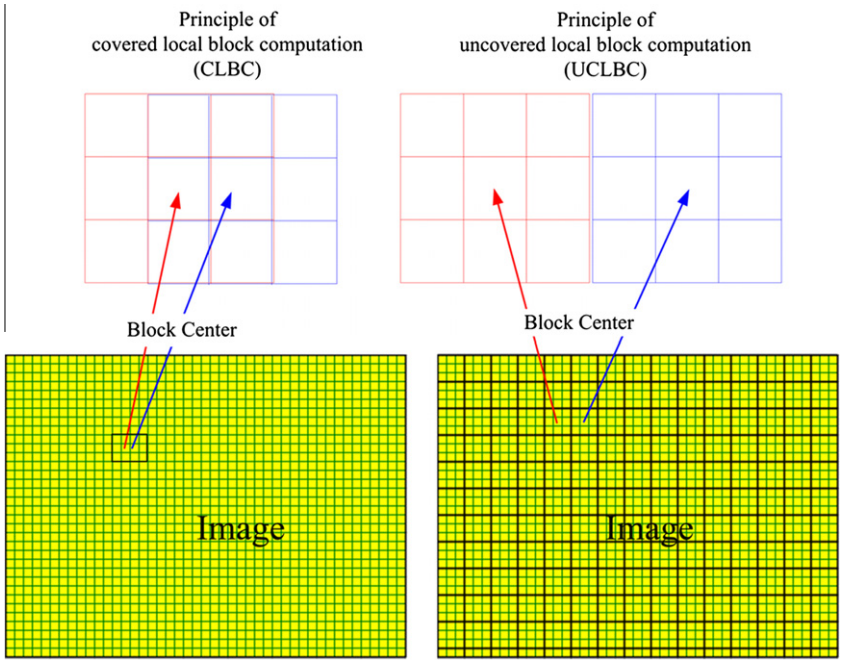


**Fig. 6.** The principle of covered local block and uncovered local block with a block size of $3 \times 3$.

**Table 2**
Number of high complexity operators used in each fusion metric.

| | Qu et al. [2] | Hossny et al. [3] | Xydeas and Petrovic [4] | Piella [5] | Cvejic et al. [6] | Yang et al. [7] | Chen and Varshney [9] | Chen and Blum [10] | Our method [21] |
|---|---|---|---|---|---|---|---|---|---|
| *N* | – | – | 6 | 22 | 8 | 9 | 7 | 7 | 7.9 |

(2) Compute *WFQ* utilizing *SSIM* metric. Because one *SSIM* can be calculated by 5 convolution operations, the algorithm requires 20 convolution operators for 4 *SSIM*.
(3) Subtracting 4 repeating computations (for fusion image) in *SSIM* computation, the total number of high complexity operators is 22.

For Blum's metric [10]

(1) Filter image by *CSF* (in spectrum space). This operation is equivalent to 1 high complexity operation.
(2) Local band contrast computation (for each image) requires 6 convolution operations (2 convolutions $\times$ 3 images).
(3) Because the remaining operations are not high complexity operators, the total number of high complexity operators is 7.

For our metric.
(1) The image is filtered into 4 sub-bands and downsampled by 2 with sub-band changes. Moreover, there are 6 convolution operations in each band.
(2) Because downsampling decreases the image area, the total number of high complexity operators is $6 \times (1 + 1/4 + 1/16 + 1/64)$.

Table 2 shows the number of high complexity operators used in each fusion metric. Note that this idea could not be used to analyze Qu and Hossny's method because it is not an *LBC*-based algorithm. It is obvious that the computational burden of *VIFF* is lower than the average because of the downsampling operation used in our metric. Considering the high predictive performance of our metric, this execution time can be acceptable.

Although this method is not a rigorous form of complexity analysis and only gives an approximate estimation of the computational burden of fusion metrics, we believe it is still meaningful and effective for comparing time complexity for most fusion metrics. Because this method relies on the fact that *LBC* is widely used in image processing, we believe it could be easily extended to time complexity analysis of other *LBC*-based algorithms.

## 5. Conclusions

A new image fusion assessment metric, *VIFF*, is presented in this paper. *VIFF* first decomposes the source and fused images. Then, *VIFF* utilizes the models in *VIF* (*GSM* model, Distortion model and *HVS* model) to capture visual information from the two source-fused pairs. With the help of an effective visual information index, *VIFF* measures the effective visual information of the fusion in all blocks in each sub-band. Finally, the assessment result is calculated by integrating all the information in each sub-band. To fairly assessment the metric's performance, an experiment was performed on Petrovic's subjective test database. Experimental results show that *VIFF* demonstrates better predictive performance. Moreover, according to our analysis, *VIFF* has lower computational complexity than other conventional fusion metrics. We believe that our work might help in the construction of more reliable fusion metrics in the future.

## Acknowledgements

The authors thank the editors and anonymous reviewers for their comments and suggestions. This work was supported jointly by the National Natural Science Foundation of China (61004088),the Key Foundation for Basic Research from Science and Technology Commission of Shanghai (09JC1408000) and the National Basic Research Program (973 Program: 6131010306).

## References

[1] M.I. Smith, J.P. Heather, Review of image fusion technology in 2005, in: Proceedings of SPIE, 2005, pp. 29–45.
[2] G.H. Qu, D.L. Zhang, P.F. Yan, Information measure for performance of image fusion, Electronics Letters (2002) 313–315.
[3] M. Hossny, S. Nahavandi, D. Creighton, Comments on 'Information measure for performance of image fusion', Electronics Letters (2008) 1066–1067.
[4] C.S. Xydeas, V. Petrovic, Objective image fusion performance measure, Electronics Letters (2000) 308–309.
[5] G. Piella, A new quality metric for image fusion, in: Proceedings of International Conference on Image Processing, 2003, pp. 173–176.
[6] N. Cvejic, A. Łoza, D. Bull, N. Canagarajah, A similarity metric for assessment of image fusion algorithms, in: International Journal of Signal Processing, 2005, pp. 178–182.
[7] C. Yang, J.Q. Zhang, X.R. Wang, X. Liu, A novel similarity based quality metric for image fusion, Information Fusion 9 (2008) 156–160.
[8] Z. Wang, A.C. Bovik, Image quality assessment: from error visibility to structural similarity, IEEE Transactions on Image Processing (2004) 600–612.
[9] H. Chen, P.K. Varshney, A human perception inspired quality metric for image fusion based on regional information, Information Fusion 8 (2007) 193–207.
[10] Y. Chen, R.S. Blum, A new automated quality assessment algorithm for image fusion, Image and Vision Computing (2008) 1421–1432.
[11] V. Petrovic, Subjective tests for image fusion evaluation and objective metric validation, Information Fusion 8 (2007) 208–216.
[12] H.R. Sheikh, A.C. Bovik, G. de Veciana, An information fidelity criterion for image quality assessment using natural scene statistics, IEEE Transactions on Image Processing (2005) 2117–2128.
[13] H.R. Sheikh, A.C. Bovik, Image information and visual quality, IEEE Transactions on Image Processing (2006) 430–444.
[14] M.J. Wainwright, E.P. Simoncelli, Scale mixtures of Gaussians and the statistics of natural images, in: Advances in Neural Information Processing Systems, MIT Press, 2000, pp. 855–861.
[15] M.J. Wainwright, E.P. Simoncelli, A.S. Willsky, Random cascades on wavelet trees and their use in analyzing and modeling natural images, Applied and Computational Harmonic Analysis (2001) 89–123.
[16] H.R. Sheikh, M.F. Sabir, A.C. Bovik, Pixel domain version of VIF. <http://live.ece.utexas.edu/research/quality/vifp_release.zip>.
[17] P. Corriveau, A. Webster, Final report from the VQEG on the validation of objective models of video quality assessment, 2000. <http://www.its.bldrdoc.gov/vqeg/>.
[18] A.M. Rohaly et al., Video quality experts group: current results and future directions, visual communications and image processing, in: Proceedings of SPIE, 2000, pp. 742–753.
[19] A. Srivastava, A.B. Lee, E.P. Simoncelli, S.C. Zhu, On advances in statistical modeling of natural images, Journal of Mathematical Imaging and Vision (2003) 17–33.
[20] ITU- Recommendation, Methodology for subjective assessment of the quality of television pictures, BT.500-10, 2000.
[21] Y. Han, image fusion metric. <http://hansy.weebly.com/image-fusion-metric.html>.