TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 05/16 pp703-709
Volume 14, Number 6, December 2009

# Objective Image Fusion Quality Evaluation Using Structural Similarity*

ZHENG Youzhi (郑有志), QIN Zheng (覃 征)**

**Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China**

**Abstract:** Objective evaluations of fused images are important in comparing the performance of different image fusion algorithms. This paper describes a structural similarity metric that does not use a reference image for image fusion evaluations. The metric is based on the universal image quality index and addresses not only the similarities between the input images and the fused image, but also the similarities among the input images. The evaluation process distinguishes between complementary information and redundant information using similarities among the input images. The metric uses the information classification to estimate how much structural similarity is preserved in the fused image. Tests demonstrate that the metric correlates well with subjective evaluations of the fused images.

**Key words:** image quality evaluation; image fusion; objective quality evaluation; structural similarity

## Introduction

Numerous imaging sensors have been developed in many fields such as military applications, medical diagnostics, machine vision, and remote sensing, with image fusion attracting much attention in image processing in recent years[1]. Image fusion combines information for a scene from two or more images from identical or different types of sensors to generate a more precise, comprehensive, and reliable image description or interpretation of that scene (usually a composite image). The objective of the image fusion is to reduce uncertainty and fully utilize complementary and redundant information in the original images. There have been many studies of image fusion[2,3], with the image quality evaluation criterion playing an important role in image fusion. Efficient performance evaluations of fusion schemes are needed to improve the development of image fusion algorithms. Also, image quality evaluations will improve the fusion performance by optimizing the image fusion algorithm parameters.

Existing image quality evaluation methods can be divided into subjective and objective evaluations[4]. In practice, human evaluations (subjective evaluation) of the fused image are of fundamental importance in many applications. However, subjective evaluations are inconvenient, time consuming, and expensive, and the test conditions cannot be guaranteed to be exactly the same. Moreover, subjective evaluations cannot be embedded into image fusion schemes to optimize the fusion algorithm parameters. Thus, objective evaluation methods are needed to compare "good" or "bad" fused images. Many image quality evaluations in the literature use an ideal fused image as a reference for comparison with the image fusion results. The mean squared error (MSE) and peak signal to noise ratio (PSNR)-based metrics were widely used for these comparisons. However, ideal fusion images are not available to most real world applications.

Therefore, objective quality evaluation methods have been developed that do not need a reference

Received: 2008-09-22; revised: 2009-05-25
* Supported by the National Natural Science Foundation of China (No. 60673024)
** To whom correspondence should be addressed.
E-mail: qingzh@tsinghua.edu.cn; Tel: 86-10-62795399

image[5-10]. A mutual information (MI) metric was used to evaluate fusion performance by Qu et al.[6,7] Xydeas and Petrovic[8] evaluated the fusion performance by calculating the relative amount of edge information transferred from the input images to the fused image. Recently, an image quality index based on the structural metric proposed by Wang and Bovik[9] was improved for image fusion assessment by Piella and Heijmans[10] into a pixel by pixel or region by region method, giving weighted averages of the similarities between the fused image and each of the source images. Cvejić et al.[11] also improved the structural metric by incorporating local measurements to estimate the amount of important information in the source images that can be represented in the fused image. However, none of these objective quality evaluations identify regions which contain complementary and redundant information in the original images.

This paper presents a structural similarity quality metric for image fusion which separately treats complementary and redundant regions in the original images. This objective quality evaluation also takes into account the amount of important information in the source images that can be transferred into the fused image. Tests indicate that the metric more effectively correlates the image information with the subjective quality measurement.

# 1 Background

Wang and Bovik[9] proposed the universal image quality index (UIQI) which does not use traditional error summation methods. The UIQI metric measures structural distortions by modeling any image distortion as a combination of the loss of correlation, luminance distortion, and contrast distortion.

Let $x = \{x_i | i = 1, 2, \ldots, N\}$ and $y = \{y_i | i = 1, 2, \ldots, N\}$ be the original and test image signals. The UIQI is then defined as

$$Q_{\mathrm{I}} = \frac{4\sigma_{xy}\bar{x}\,\bar{y}}{(\sigma_x^2 + \sigma_y^2)\cdot(\bar{x}^2 + \bar{y}^2)} = \frac{\sigma_{xy}}{\sigma_x\sigma_y}\cdot\frac{2\bar{x}\,\bar{y}}{\bar{x}^2 + \bar{y}^2}\cdot\frac{2\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2} \tag{1}$$

where

$$\bar{x} = E(x) = \frac{1}{N}\sum_{i=1}^{N} x_i\ ,\quad \bar{y} = E(y) = \frac{1}{N}\sum_{i=1}^{N} y_i\ ,$$

$$\sigma_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2\ ,\quad \sigma_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2\ ,$$

$$\sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})\ .$$

The $Q_{\mathrm{I}}$ range is [−1, 1]. The best value 1 is achieved if and only if $y_i = x_i$ for all $i$ = 1, 2, … , $N$. The lowest value of −1 occurs when $y_i = 2E(x) - x_i$ for all $i$ =1, 2, … , $N$. In Eq. (1), the first term is the correlation coefficient between $x$ and $y$ which has a range of [−1, 1]. The second term has a range of [0, 1] and measures the similarities of the mean luminances of $x$ and $y$. The third term, which also has a range of [0, 1], measures the similarities of the images contrasts.

Since image signals are generally non-stationary, the image quality index $Q_{\mathrm{I}}$ is best measured over local regions and then combined into a single measure. Wang and Bovik[9] used a sliding window approach. The fixed size (with $n$ pixels) window starts from the top-left corner of the image and moves pixel by pixel horizontally and vertically through all the rows and columns of the image until reaching the bottom-right corner. The local quality index $Q_{\mathrm{I}}(x, y|w)$ is computed for the pixels within the sliding window $w$ for each window. Finally, the overall image quality index $Q_{\mathrm{I}}(x, y)$ is computed by averaging all the local quality indices,

$$Q_{\mathrm{I}}(x, y) = \frac{1}{|W|}\sum_{w\in W} Q_{\mathrm{I}}(x, y | w) \tag{2}$$

where $W$ is the family of all sliding windows and $|W|$ is the number of windows.

To apply the UIQI for image fusion evaluation, Piella and Heijmans[10] defined the saliency of two original images within the sliding window as $s(x|w)$ and $s(y|w)$. This metric reflects the local relevance of an image within window $w$ as a function of the contrast, sharpness, or entropy. Given the local saliencies $s(x|w)$ and $s(y|w)$ of the two input images, they compute a local weight $\lambda(w)$ between 0 and 1 that indicates the relative importance of image $x$ compared to image $y$.

$$\lambda(w) = \frac{s(x|w)}{s(x|w) + s(y|w)} \tag{3}$$

The objective quality evaluation that does not use a reference image is defined then as[10]

$$Q_{\mathrm{P}} = \sum_{w\in W} c(w)(\lambda(w)Q_{\mathrm{I}}(x, f|w) + (1 - \lambda(w))Q_{\mathrm{I}}(y, f|w)) \tag{4}$$

where $x$ and $y$ are the input images, $f$ represents the fused image, and $c(w)$ is the overall saliency of window $w$ defined as

$$c(w)=\mathrm{Max}(s(x|w), s(y|w))/(\sum_{w'\in W}\mathrm{Max}(s(x|w'), s(y|w'))) \tag{5}$$

To compute the metric in Eq. (4), $\lambda$ in Eq. (3) is computed using $s(x|w)$ and $s(y|w)$ which are the variances (or the average in edge images) of images $x$ and $y$ within window $w$. Therefore, the metric does not consider the relationship between the input images and the final fused image. In addition, each time the metric is calculated, an "edge image" has to be derived from the input images, which adds significantly to the computational complexity of the metric. Cvejić et al.[11] proposed a performance measure that takes into account the similarities between the input image block and the fused image block within the sliding window:

$$\begin{aligned} Q_{\mathrm{N}} &= \frac{1}{|W|}\sum_{w\in W}\beta(x,y,f\,|\,w)Q_{\mathrm{I}}(x,f\,|\,w)+ \\ &\quad (1-\beta(x,y,f\,|\,w))Q_{\mathrm{I}}(y,f\,|\,w)= \\ &\quad \frac{1}{|W|}\sum_{w\in W}\beta(x,y,f\,|\,w)(Q_{\mathrm{I}}(x,f\,|\,w)- \\ &\quad Q_{\mathrm{I}}(y,f\,|\,w))+Q_{\mathrm{I}}(y,f\,|\,w) \end{aligned} \tag{6}$$

where

$$\beta(x,y,f\,|\,w)= \begin{cases} 0, & \text{if } \sigma_{xf}/(\sigma_{xf}+\sigma_{yf})<0; \\ \sigma_{xf}/(\sigma_{xf}+\sigma_{yf}), & \text{if } 0\leqslant\sigma_{xf}/(\sigma_{xf}+\sigma_{yf})\leqslant 1; \\ 1, & \text{if } \sigma_{xf}/(\sigma_{xf}+\sigma_{yf})>1 \end{cases} \tag{7}$$

$$\sigma_{uv}=\frac{1}{N-1}\sum_{i=1}^{N}(u_i-\bar{u})(v_i-\bar{v})\,.$$

## 2 Structural Similarity Metric

The parameter $\beta$ in Eq. (6) takes into account the similarity between the input images and the fused image. However, there is no clear measure of the similarity among input images. The aim of image fusion is to integrate complementary and redundant information from multiple images to create a composite image that contains a better description of the scene than any of the individual source images. Therefore, a good fusion performance evaluation should correctly distinguish the regions which contain complementary and redundant information in the original images, and should estimate how much information is preserved in the fused image.

Good empirical fusion rules used in many applications are the averaging rule for redundant information and the selecting rule for complementary information[12]. If the redundant information is treated in the same way as the complementary information in evaluating fusion schemes, schemes that successfully eliminate conflicting information will rank below those that do not, which is obviously not correct.

The current structural similarity metric makes use of a local structural matching measure between the source images. The basic idea is to use information classification on all the source images for evaluation of the image fusion. This objective quality evaluation is also based on the UIQI, and uses the sliding window approach.

Local structural matching measures $\theta(x, y|w)$, $\theta(x, f|w)$, and $\theta(y, f|w)$, are calculated for each sliding window ($x$ and $y$ are the input images, $f$ is the fused image, and $w$ is the analysis window). The measure is defined as

$$\theta(u,v\,|\,w)=\frac{1}{N}\sum_{i=1}^{N}\frac{2u_iv_i}{u_i^2+v_i^2} \tag{8}$$

which uses the local structural similarity between the input images $\theta(x, y|w)$ as the detection criterion. Sliding windows whose $\theta(x, y|w)$ are equal to or larger than a given threshold $T$ (set as 0.8 here) are assumed to contain redundant information in the analysis window. Sliding windows whose $\theta(x, y|w)$ are less than the given threshold are assumed to contain complementary information in the window.

The local quality metric for each sliding window is defined according to the window classification, as shown in Eq. (9). The maximum UIQI metric is selected for a complementary window and the weighted average of the UIQI metric is used for a redundant window.

$$Q_{\mathrm{S}}(x,y\,|\,w)= \begin{cases} \mathrm{Max}(Q_{\mathrm{I}}(x,f\,|\,w), Q_{\mathrm{I}}(y,f\,|\,w)), & \text{if } \theta(x,y\,|\,w)<T; \\ \dfrac{\theta(x,f\,|\,w)}{\theta(x,f\,|\,w)+\theta(y,f\,|\,w)}Q_{\mathrm{I}}(x,f\,|\,w)+\dfrac{\theta(y,f\,|\,w)}{\theta(x,f\,|\,w)+\theta(y,f\,|\,w)}Q_{\mathrm{I}}(y,f\,|\,w), & \text{if } \theta(x,y\,|\,w)\geqslant T \end{cases} \tag{9}$$

The local quality metric for each sliding window is dependent not only on the similarities between the input images and the fused image but also on the similarities among the input images. In the end, the overall image quality evaluation $Q_{\mathrm{S}}(x, y)$ is computed by averaging all the local quality metrics:

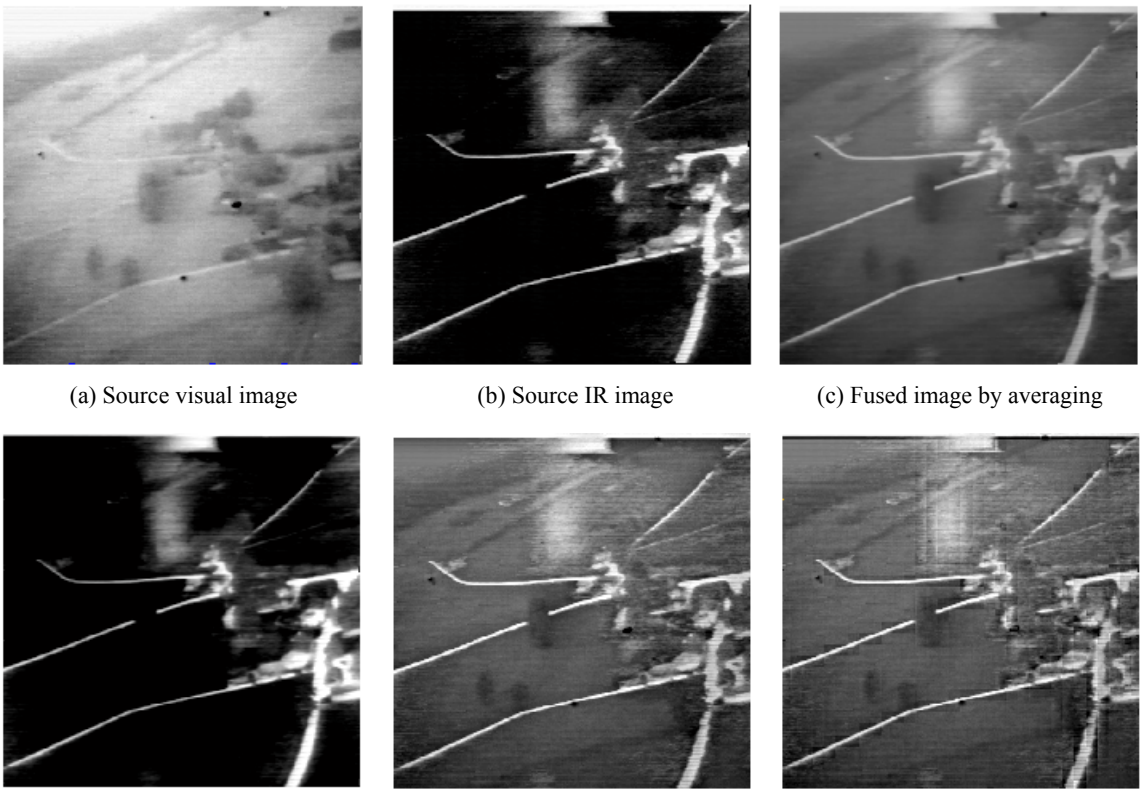
$$Q_S(x,y)=\frac{1}{|W|}\sum_{w\in W}Q_S(x,y\,|\,w) \tag{10}$$

where $W$ and $|W|$ again represent the family of all sliding windows and the number of windows. $Q_S(x, y)$ closer to 1 indicates higher fused image quality.

## 3 Test Results

The fusion quality measure has been used to evaluate four image fusion methods with comparisons to other standard objective metrics. The first image fusion method is a simple averaging of the two source images, while the second is based on a principal component analysis (PCA) algorithm. The multi-resolution (MR) image fusion Laplacian pyramid[13] and Haar wavelet transform (HWT) methods are also used with four-level decomposition in each case. The coefficients of the MR decomposition were selected as the maximum absolute value for the detailed images at each position, as well as the mean of the approximation images. The four standard objective metrics used for comparison were the MI[6], Piella's metric $Q_P$[10], Cvejić's metric $Q_N$[11], and Xydeas' metric[8]. The sliding window size was set as 8×8.

The first test used a military scouting application, where the source images were produced by a visual sensor (Fig. 1a) and an infra-red (IR) sensor (Fig. 1b). In this case, the complementary information regions and redundant information regions are difficult to distinguish. Figures 1c-1f show the four fused images obtained using the four fusion schemes. The Laplacian method outperforms the three other methods. The averaging and PCA methods lost many details. The HWT method distorted the images while preserving some details. Table 1 lists the four objective metrics for the fused images in Fig. 1. The results show that the current structural similarity quality metric agrees with the subjective evaluation and three of the other standard structural metrics that the Laplacian pyramid image is the best of the fused images (the MI method selected the PCA image as the best).



(a) Source visual image (b) Source IR image (c) Fused image by averaging

(d) Fused image by PCA (e) Fused image by Laplacian pyramid (f) Fused image by HWT

**Fig. 1 Fusion of visual and IR images**

**Table 1 Comparison of various metrics for the fused images in Fig. 1**

| | MI | $Q_P$ | $Q_N$ | Xydeas | Current |
|---|---|---|---|---|---|
| Averaging | 2.664 | 0.489 | 0.424 | 0.167 | 0.362 |
| PCA | 3.290 | 0.461 | 0.228 | 0.095 | 0.261 |
| Laplacian | 1.985 | 0.644 | 0.643 | 0.418 | 0.563 |
| HWT | 1.945 | 0.607 | 0.595 | 0.402 | 0.524 |

Two out-of-focus images and the fused images are shown in Fig. 2. In this case, the information in the two source images is complementary (Fig. 2a with focus on the right and Fig. 2b with focus on the left). Subjectively, the first two methods are obviously poorer than the last two methods. However, the subjective evaluation cannot easily distinguish whether the Laplacian pyramid or the HWT result is better or whether the averaging or the PCA result is worse. The various objective metrics for the fused images in Fig. 2 are listed in Table 2. The present metric shows that the Laplacian pyramid and HWT images are much better than the averaging and PCA results as expected. Although the current metric selects the HWT result as the best and the averaging as the worst, the difference between the Laplacian pyramid and HWT results, as well as the difference between the averaging and PCA results are quite small. Thus, the current metric gives the best agreement with the subjective evaluation among all these metrics.

(a) Image with focus on the right (b) Image with focus on the left (c) Fused image by averaging

(d) Fused image by PCA (e) Fused image by Laplacian pyramid (f) Fused image by HWT

**Fig. 2 Fusion of out-of-focus images**

**Table 2 Comparison of various metrics for the fused images in Fig. 2**

| | MI | $Q_P$ | $Q_N$ | Xydeas | Current |
|---|---|---|---|---|---|
| Averaging | 6.807 | 0.812 | 0.734 | 0.483 | 0.683 |
| PCA | 6.807 | 0.812 | 0.734 | 0.480 | 0.684 |
| Laplacian | 6.709 | 0.850 | 0.790 | 0.746 | 0.840 |
| HWT | 6.252 | 0.834 | 0.811 | 0.686 | 0.851 |

The third pair of test images is Wang Zhaojun, one of the four ancient Chinese beauties. The test images shown in Fig. 3 were artificially created by blurring the original image using Gaussian noise. The images are complementary in the sense that the blurring occurs at different locations. For comparison, the PSNR between the reference image (the original Wang Zhaojun image) and each of the generated fused images was calculated to evaluate the fused image, where the PSNR is defined as

$$\mathrm{PSNR} = 10\lg\frac{255\times 255}{\mathrm{RMSE}^2},$$
$$\mathrm{RMSE} = \sqrt{\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}[R(i,j)-F(i,j)]^2} \qquad (11)$$

The various metrics for the fused images in Fig. 3 are listed in Table 3, which shows that the current metric gives the same ranking of the fusion methods as the PSNR metric. In addition, the subjective results also prove that our proposed metric gives a better evaluation performance.

(a) Source image *x* (b) Source image *y* (c) Fused image by averaging

(d) Fused image by PCA (e) Fused image by Laplacian pyramid (f) Fused image by HWT

**Fig. 3 Fusion of blurred and Gaussian noise augmented images**

**Table 3 Comparison of various metrics for the fused images in Fig. 3**

| | MI | $Q_P$ | $Q_N$ | Xydeas | Current | PNSR |
|---|---|---|---|---|---|---|
| Averaging | 6.174 | 0.820 | 0.727 | 0.601 | 0.732 | 12.500 |
| PCA | 6.166 | 0.821 | 0.730 | 0.593 | 0.741 | 12.576 |
| Laplacian | 6.561 | 0.858 | 0.691 | 0.819 | 0.782 | 12.727 |
| HWT | 6.294 | 0.856 | 0.725 | 0.771 | 0.743 | 12.653 |

## 4 Conclusions

A structural similarity quality metric was developed as an objective non-reference performance evaluation for an image fusion method which takes into account not only the similarities between the input images and the fused an image, but also the similarities among the input images. Comparisons with other standard objective quality metrics show that this metric correlates well with subjective quality evaluation of the fused images, especially for input images where the complementary information and the redundant information can be well distinguished.

There are several areas in which our quality metrics can be improved or extended. Firstly, the criterion for distinguishing between complementary and redundant information can be defined better. Secondly, other visual mechanisms used in human visual system or quality metrics can also be embedded into the metric. Finally, the quality metrics can evaluate regions of the image using image segmentation rather than fixed square windows.

## References


[1] Piella G. Adaptive wavelets and their applications to image fusion and compression [Dissertation]. CWI University of Amsterdam, Holland, 2003.

[2] Goshtasby A A, Nikolov S. Image fusion: Advances in the state of the art. *Information Fusion*, 2007, **8**(2): 114-118.

[3] Moira I S, Jamie P H. Review of image fusion technology in 2005. In: The International Society for Optical Engineering (SPIE). Bellingham, WA, USA, 2005: 29-45.

[4] Petrović V. Subjective tests for image fusion evaluation and objective metric validation. *Information Fusion*, 2007, **8**(2): 208-216.

[5] Cvejić N, Bull D, Canegarajah C A. New metric for multimodal image sensor fusion. *Electronics Letters*, 2007, **43**(2), 95-96.

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

[7] Hossny M, Nahavandi S, Creighton D. Comments on "Information measure for performance of image fusion". *Electronics Letters*, 2008, **44**(18): 1066-1067.

[8] Xydeas C S, Petrovic V. Objective image fusion performance measure. *Electronics Letters*, 2000, **36**(4): 308-309.

[9] Wang Z, Bovik A C. A universal image quality index. *IEEE Signal Processing Letters*, 2002, **9**(3): 81-84.

[10] Piella G, Heijmans H. A new quality metric for image fusion. In: Proc. IEEE Int. Conf. on Image Processing. Barcelona, Spain, 2003: 173-176.

[11] Cvejić N, Łoza A, Bull D, et al. A similarity metric for assessment of image fusion algorithms. *International Journal of Signal Processing*, 2005, **2**(3): 178-182.

[12] Zheng Youzhi, Hou Xiaodong, Bian Tiantian, et al. Effective image fusion rules of multi-scale image decomposition. In: International Symposium on Image and Signal Processing and Analysis. Istanbul, Turkey, 2007: 362-366.

[13] Burt P J, Adelson E H. The Laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 1983, **31**(4): 532-540.