International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:1, No:7, 2007

A Novel Metric for Performance Evaluation of Image Fusion Algorithms

Nedeljko Cvejic, Artur Łoza, David Bull, and Nishan Canagarajah

***Abstract*—**In this paper, we present a novel objective non-reference performance assessment algorithm for image fusion. It takes into account local measurements to estimate how well the important information in the source images is represented by the fused image. The metric is based on the Universal Image Quality Index and uses the similarity between blocks of pixels in the input images and the fused image as the weighting factors for the metrics. Experimental results confirm that the values of the proposed metrics correlate well with the subjective quality of the fused images, giving a significant improvement over standard measures based on mean squared error and mutual information.

***Keywords*—**Fusion performance measures, image fusion, non-reference quality measures, objective quality measures.

I. Introduction

IMAGE and video fusion is emerging as a vital technology in many military, surveillance and medical applications. It is a subarea of the more general topic of data fusion, dealing with image and video data [1,2]. The ability to combine complementary information from a range of distributed sensors with different modalities can be used to provide enhanced performance for visualization, detection or classification tasks. Multi-sensor data often present complementary information about the scene or object of interest, and thus image fusion provides an effective method for comparison and analysis of such data. There are several benefits of multi-sensor image fusion: wider spatial and temporal coverage, extended range of operation, decreased uncertainty, improved reliability and increased robustness of the system performance.

In several application scenarios, image fusion is only an introductory stage to another task, e.g. human monitoring. Therefore, the performance of the fusion algorithm must be measured in terms of improvement in the following tasks. For example, in classification systems, the common evaluation measure is the number of the correct classifications. This system evaluation requires that the 'true' correct classifications are known. However, in experimental setups the ground-truth data might not be available.

Authors are with the Centre for Communications Research, University of Bristol, Merchant Venturers Building, Woodland Road, Bristol BS8 1UB, United Kingdom, (Corresponding author phone: +44 117 331 5102; fax: +44 117 954 5206, e-mail: n.cvejic@bristol.ac.uk).

This work has been funded by the UK Ministry of Defence Data and Information Fusion Defence Technology Centre.

In many applications the human perception of the fused image is of fundamental importance and as a result the fusion results are mostly evaluated by subjective criteria [3,4]. Objective image fusion performance evaluation is a tedious task due to different application requirements and the lack of a clearly defined ground-truth. Various fusion algorithms presented in the literature [5] have been evaluated objectively by constructing an "ideal" fused image and using it as a reference for comparison with the experimental results [6,7]. Mean squared error (MSE) based metrics were widely used for these comparisons. Several objective performance measures for image fusion have been proposed where the knowledge of ground-truth is not assumed. In [8], authors used the mutual information as a parameter for evaluation of the fusion performance. Xydeas and Petrovic [9] proposed a metric that evaluates the relative amount of edge information that is transferred from the input images to the fused image.

In this paper, we present a novel objective non-reference quality assessment algorithm for image fusion. It takes into account local measurements to estimate how well the important information in the source images is represented by the fused image, while minimizing the number of artefacts or the amount of distortion that could interfere with interpretation. Our quality measures are based on an image quality index proposed by Wang and Bovik [10].

II. Definition of the Universal Image Quality Index

The measure that was used as the basis for our objective performance evaluation of image fusion is the Universal Image Quality Index (UIQI) [10]. The authors compared the proposed quality index to the standard MSE objective quality measure and the main conclusion was that their new index outperforms the MSE, due to the UIQI's ability in measuring structural distortions [10].

Let $X=\{x_i|i=1,2,\ldots,N\}$ and $Y=\{y_i|i=1,2,\ldots,N\}$ be the original and the test image signals, respectively. The proposed quality index is defined as [10]:

$$Q=\frac{4\sigma_{xy}\bar{x}\cdot\bar{y}}{\left(\sigma_x^2+\sigma_y^2\right)\left[\left(\bar{x}\right)^2+\left(\bar{y}\right)^2\right]} \tag{1}$$

where

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

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

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

The dynamic range of Q 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 $y_i=2\bar{x}-x_i$ for all $i=1,2,...N$. This quality index models image distortions as a combination of three different factors: loss of correlation, luminance distortion and contrast distortion. In order to make this more understandable, the definition of Q can be rewritten as a product of three components:

$$Q=\frac{\sigma_{xy}}{\sigma_x\sigma_y}\frac{2\bar{x}\cdot\bar{y}}{(\bar{x})^2+(\bar{y})^2}\frac{2\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2} \quad (3)$$

The first component is the correlation coefficient between X and Y and its dynamic range is [-1,1]. The best value 1 is obtained when $y_i=ax_i+b$ for all $i=1,2,...N$, where a and b are constants and $a>0$. Even if X and Y are linearly related, there still might be relative distortions between them and these are evaluated in the second and third component. The second component with a value range of [0,1] measures how close the mean luminance is between X and Y. It equals 1 if and only if $\bar{x}=\bar{y}$. σ_x and σ_y can be viewed as an estimate of the contrast of X and Y, so the third component measures how similar the contrasts of the images are. The range of values for the third component is also [0,1], where the best value 1 is achieved if and only if $\sigma_x=\sigma_y$.

Since images are generally non-stationary signals, it is appropriate to measure Q_0 over local regions and then combine the different results into a single measure Q. In [10] the authors propose to use a sliding window: starting from the top-left corner of the two images X, Y, a sliding window of a fixed size block by block over the entire image until the bottom-right corner is reached. For each window w the local quality index $Q_0(X,Y|w)$ is computed for the pixels within the sliding window w. Finally, the overall image quality index Q is computed by averaging all local quality indices:

$$Q(x,y)=\frac{1}{|W|}\sum_{w\in W}Q_0(a,b\,|\,w) \quad (4)$$

where W is the family of all windows and $|W|$ is the cardinality of W. Wang and Bovik [10] have compared (under several types of distortions) their quality index with existing image measures such as MSE as well as with subjective evaluations. The tested images were distorted by: additive white Gaussian noise, blurring, contrast stretching, JPEG compression, salt and pepper noise, mean shift and multiplicative noise. The main conclusion was that UIQI outperforms the MSE, which due to the index's ability of measuring structural distortions, in contrast to the MSE which is highly sensitive to the energy of errors.

In order to apply the UIQI for image fusion evaluation, Piella and Heijmans [11] introduce salient information to the metric.

$$Q_p(X,Y,F)=\sum_{w\in W}c(w)\left(\lambda Q(X,F\,|\,w)+(1-\lambda)Q(Y,F\,|\,w)\right) \quad (5)$$

where X and Y are the input images, F is the fused image, $c(w)$ is the overall saliency of a window and λ is defined as:

$$\lambda=\frac{s(X\,|\,w)}{s(X\,|\,w)+s(Y\,|\,w)}. \quad (6)$$

λ should reflect the relative importance of image X compared to image Y within the window w. $s(X|w)$ denotes saliency of image X in window w. It should reflect the local relevance of image X within the window w, and it may depend on e.g. contrast, sharpness, or entropy. As with the previous metrics, this metric does not require a ground-truth or reference image. Finally, to take into account some aspect of the human visual system (HVS) which is the relevance of edge information, the same measure is computed with the 'edge images' (X', Y' and F') instead of the grey-scale images X, Y and F.

$$Q_E(X,Y,F)=Q_p(X,Y,F)^{1-\alpha}Q_p(X',Y'F')^{\alpha} \quad (7)$$

where α is a parameter that expresses the contribution of the edge images compared to the original images.

III. Proposed Image Fusion Performance Metrics

In the computation of Piella's metric parameter λ in equation (4) is computed with $s(X|w)$ and $s(Y|w)$ being the variance (or the average in the edge images) of images X and Y within window w, respectively. Therefore, there is no clear measure of how similar each input image is to the final fused image. 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. In addition, the metrics calculated and presented in [11] are only for one window size (8x8). The window size has a significant influence on this fusion performance measure, as the main weighting factor is the ratio of the variances of the input images which tend to vary significantly with the window size.

We propose a novel fusion performance measure that takes into account the similarity between the input image block and the fused image block within the same spatial position. It is defined as:

$$\begin{aligned}Q_b(X,Y,F)&=\sum_{w\in W}sim(X,Y,F\,|\,w)Q(X,F\,|\,w)+(1-sim(X,Y,F\,|\,w))Q(Y,F\,|\,w)\\&=\sum_{w\in W}sim(X,Y,F\,|\,w)\cdot(Q(X,F\,|\,w)-Q(Y,F\,|\,w))+Q(Y,F\,|\,w)\end{aligned} \quad (8)$$

where X and Y are the input images, F is the fused image, w is the analysis window and W is the family of all windows. We define $sim(X,Y,F|w)$ as:

$$sim(X,Y,F\,|\,w)=\begin{cases}0 & if \quad \frac{\sigma_{xf}}{\sigma_{xf}+\sigma_{yf}}<0\\ \frac{\sigma_{xf}}{\sigma_{xf}+\sigma_{yf}} & if \quad 0\le\frac{\sigma_{xf}}{\sigma_{xf}+\sigma_{yf}}\le 1\\ 1 & if \quad \frac{\sigma_{xf}}{\sigma_{xf}+\sigma_{yf}}>1\end{cases} \quad (9)$$

where

$$\sigma_{uv}=\frac{1}{N-1}\sum_{i=1}^{N}(u_i-\bar{u})(v_i-\bar{v}) \quad (10)$$

Each analysis window is weighted by the $sim(X,Y,F|w)$ that is dependent on the similarity in spatial domain between the input image and the fused image. The image block from two of the input images that is more similar to the fused image block is assigned a larger weighting factor used for calculation of the fusion performance metric. The impact of the less similar block is accordingly decreased. In this sense, we are able to measure more accurately the fusion performance, especially in an experimental setup where the input images are distorted versions of the ground-truth data; obtained by e.g. blurring, JPEG compression, noise addition, mean shift, etc.

The *sim(X,Y,F|w)* function is designed to have the upper limit at one, so that impact of the less significant block is completely eliminated when the other input block similarity measure equals one. Calculation of the *sim(X,Y,F|w)* function is computationally significantly less demanding, compared to the metrics proposed in [8] and [11].

IV. Experimental Results

In this section we test the proposed fusion quality measure in Eq. (8) to evaluate several multiresolution (MR) image fusion algorithms and compare it to standard objective image metrics. The MR-based image fusion approach consists of performing an MR transform on each input image and, following specific rules, combining them into a composite MR representation. The composite image is obtained by applying the inverse transform on this composite MR representation [2].

During the tests we use the simple averaging method, the ratio pyramid, Principal Component Analysis (PCA) method and the discrete wavelet transform (DWT), and in all MR cases we perform 5-level decomposition. We perform the fusion of the coefficients of the MR decomposition of each input image by selecting at each position the coefficient with a maximum absolute value, except for the coefficients from the lowest resolution where the fused coefficient equals to the mean value of the coefficients in that subband.

The first pair of test images used is the complementary pair shown in the top row of Fig. 1. The test images have been created artificially by blurring the original 'Goldhill' image of size 512x512, using Gaussian blurring with a radius of 10 pixels. The images are complementary in the sense that the blurring takes place at the complimentary horizontal strips in the first and the second image, respectively. The fused images obtained by the average method, the ratio pyramid, the PCA method and DWT domain fusion are depicted in the first and the second row, from left to right. Table 1 compares the quality of these composite images using our proposed quality measures. The first three rows correspond to the proposed fusion quality measure, as defined in Eq. (4). The rows 4 to 6 show the proposed fusion performance measure defined in Eq. (8). The proposed metrics are calculated for three window sizes: 4x4, 8x8 and 16x16 pixels, in order to examine the dependence of the metric's output values versus the analysis window size.

For comparison, we also compute the PSNR between the original 'Goldhill' image and each of the generated fused images. In 'real life' image fusion scenarios we do not have access to the original image, so the PSNR value is provided just as a reference. In addition, we have provided as references the fusion performance metric developed by Petrovic and Xydeas [8] (given in the fourth row of the Table 1-3) and the metric based on mutual information [9] (the fifth row of the Table 1-3). Petrovic and Xydeas metric measures the amount of edge information 'transferred' from the source image to the fused image in order to give an estimation of the performance of the fusion algorithm. It uses a Sobel edge operator to calculate the strength and orientation information of each pixel in the input and output images. In this method the visual information is primarily associated with the edge information, while the region information is ignored. More precisely, the results in the fifth row of Table 1 have been obtained by adding the mutual information between the composite image and each of the inputs and dividing it by the sum of the entropies of the inputs:

$$MI(X,Y,F) = \frac{I(X,F) + I(Y,F)}{H(X) + H(Y)} \tag{11}$$

where I(X,F) is the mutual information between X and F ,and H(X) the entropy of X. In this way, the measure is normalized to the range [0,1].

The following two pairs of input images are contaminated by or Gaussian additive noise (Fig.2) and Salt and Pepper (SP) noise (Fig.3). Although the additive noise can be tackled by performing hard thresholding of the parameters in the transform domain and SP noise by median filtering we did not perform denoising in order to get more balanced data for the proposed metric. The results for the noisy input images are given in the Table 2 and Table 3 for the image distorted by Gaussian additive noise and SP noise, respectively.

TABLE I
Comparison of different objective quality measures for the composite images in Fig. 1

metrics	average	ratio	PCA	DWT
Q_b (4x4)	0.7802	0.7232	0.7805	**0.8770**
Q_b (8x8)	0.7899	0.7485	0.7902	**0.8770**
Q_b (16x16)	0.8121	0.7762	0.8121	**0.8725**
Petrovic	0.3445	0.4189	0.3544	**0.6598**
MI	0.3158	**0.3312**	0.3173	0.2846
PSNR (dB)	28.27	23.92	28.23	**32.34**

TABLE II
Comparison of different objective quality measures for the composite images in Fig. 2

metrics	average	ratio	PCA	DWT
Q_b (4x4)	0.9321	0.8942	0.9327	**0.9924**
Q_b (8x8)	0.9328	0.8967	0.9333	**0.9895**
Q_b (16x16)	0.9337	0.8958	0.9342	**0.9808**
Petrovic	0.8619	0.8601	0.8626	**0.9745**
MI	0.6643	0.6054	**0.6644**	0.5338
PSNR (dB)	**17.09**	**17.09**	15.89	16.10

TABLE III
Comparison of different objective quality measures for the composite images in Fig. 3

metrics	average	ratio	PCA	DWT
Q_b (4x4)	0.8969	0.8385	0.8974	**0.9679**
Q_b (8x8)	0.8990	0.8601	0.8997	**0.9814**
Q_b (16x16)	0.9016	0.8665	0.9021	**0.9705**
Petrovic	0.7734	0.7889	0.7745	**0.9498**
MI	0.5366	0.4658	**0.5369**	0.4131
PSNR (dB)	**19.71**	17.99	**19.71**	18.38

Test results show that the DWT domain fusion visually outperform the other three schemes. It is most noticeable as,

for instance, the blurring (e.g., edges in the background and small details) and the loss of texture in the fused image obtained by the ratio pyramid and averaging. Furthermore, in the ratio-pyramid method fused image, some details of the images and background have been completely lost, and in the average composite image, the loss of contrast is very evident. These subjective visual comparisons agree with by the results obtained by the proposed metric, presented in Table 1-3. Note that the proposed metric has very similar quality measures as the Petrovic's metric and that these two metrics considerably outperform the MI measure and PSNR. It is clear from the experiments that MI metric and PSNR often assign the highest value of the fusion performance measure to the algorithm that does not perform well in the subjective terms. The values obtained from the proposed metrics correlate well to the subjective quality of the fused images, which was not achievable by the standard MI fusion performance measure and PSNR. In addition, the proposed metrics is not significantly dependent on the size of the analysis window as the difference in fusion performance does not change extensively with the variation of window size.

V. Conclusion

We present a novel objective non-reference performance assessment algorithm for image fusion. It takes into account local measurements to estimate how well the important information in the source images is represented by the fused image. Experimental results confirm that the values of the proposed metrics correlate well with the subjective quality of the fused images, giving a significant improvement over standard measures based on mean squared error and mutual information. Compared to already presented fusion performance measures [8,11], it obtains comparable results with considerably decreased computational complexity.

Further research will focus on how to select the salient points in order to optimize the fusion performance. Another extension of the work will be performance measure based on regions of the image, obtained by segmentation of the input images, rather than calculating the measure in square windows.

References

[1] H. Maitre and I. Bloch, "Image fusion", *Vistas in Astronomy*, Vol. 41, No. 3, pp. 329-335, 1997.
[2] S. Nikolov, P. Hill, D. Bull, and N. Canagarajah, "Wavelets for image fusion", *Wavelets in Signal and Image Analysis*, Kluwer, Dordrecht, The Netherlands, pp. 213-244, 2001.
[3] D. Ryan and R. Tinkler, "Night pilotage assessment of image fusion", *Proc. SPIE*, Vol. 2465, Orlando, FL, pp. 50–67, 1995.
[4] A. Toet and E. M. Franken "Perceptual evaluation of different image fusion schemes", *Displays*, Vol. 24, No. 1, pp. 25–37, 2003.
[5] G. Piella, "A general framework for multiresolution image fusion: from pixels to regions", *Information Fusion*, Vol. 9, pp. 259–280, 2003.
[6] H. Li, B. S. Manjunath, and S. K. Mitra, "Multisensor image fusion using the wavelet transform", *Graphical Models and Image Processing*, Vol. 57, No. 3, pp. 235–245, 1995.
[7] O. Rockinger, "Image sequence fusion using a shift invariant wavelet transform. *Proc. IEEE International Conference on Image Processing*, Washington, DC, pp. 288–291, 1997.
[8] C. Xydeas and V. Petrovic "Objective pixel-level image fusion performance measure", *In Proc. SPIE*, Vol. 4051, Orlando, FL, pp. 88–99, 2000
[9] G. H. Qu, D. L. Zhang, and P. F. Yan "Information measure for performance of image fusion", *Electronics Letters*, Vol. 38, No. 7, pp. 313–315, 2002.
[10] Z. Wang and A. C. Bovik. "A universal image quality index" *IEEE Signal Processing Letters*, Vol. 9, No. 3, pp. 81–84, 2002.
[11] G. Piella and H. Heijmans, "A new quality metric for image fusion" *In Proc. Int. Conf. Image Processing*, Barcelona, Spain, pp. 173-6, 2003.

Fig. 1 Fusion results, the original image blurred in stripes. Top: input image X (one half of stripes in the original image blurred, left), input image Y (other half of stripes in the original image blurred, middle), fused image F using averaging (right). Bottom: fused image F using ratio pyramid decomposition (left), fused image F using the PCA decomposition (middle), fused image F using DWT domain fusion (right)

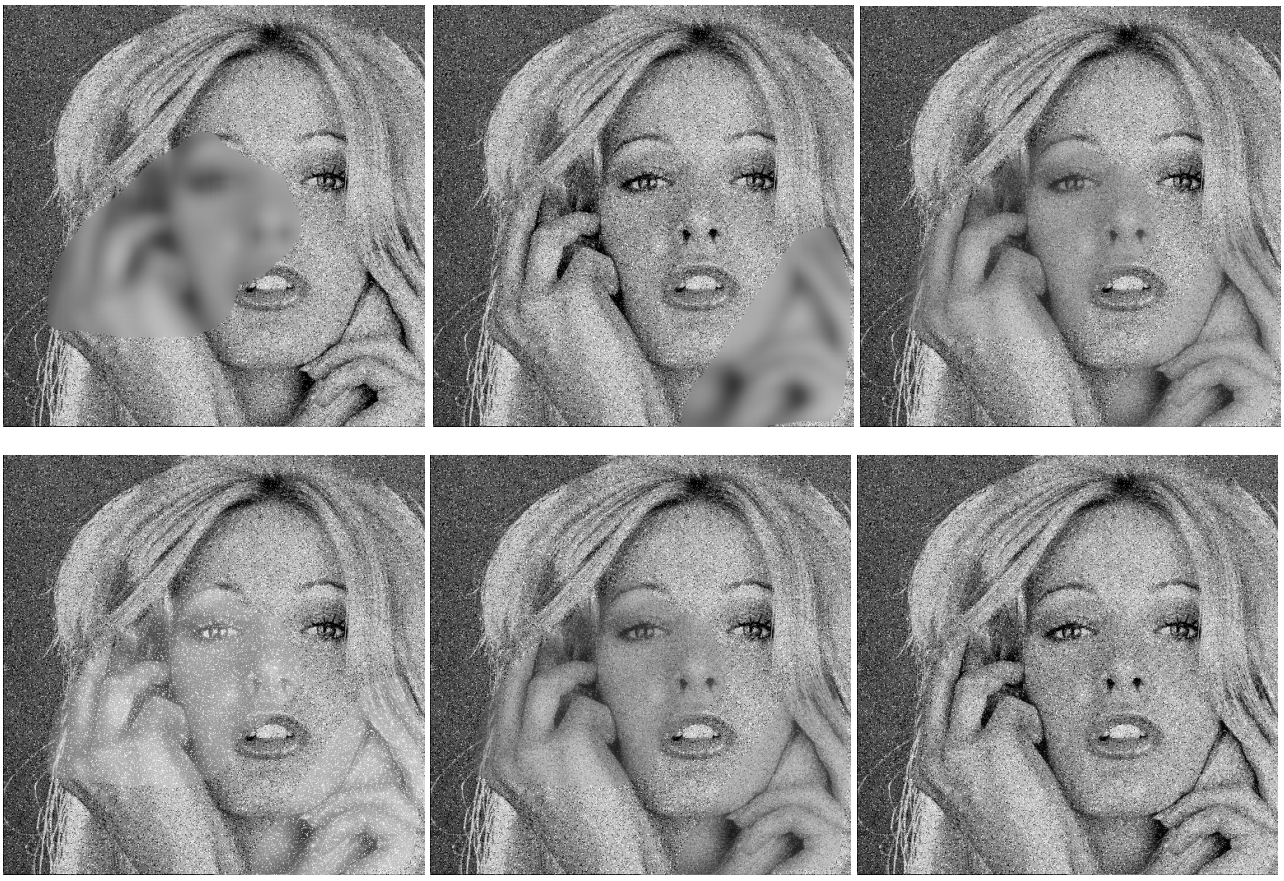

Fig. 2 Fusion results, the original image corrupted by additive Gaussian noise and partial blurring. Top: input image X (the original image with added noise and a blurred segment, left), input image Y (the original image with added noise and another blurred segment, middle), fused image F using averaging (right). Bottom: fused image F using ratio pyramid decomposition (left), fused image F using the PCA decomposition (middle), fused image F using DWT domain fusion (right)

Fig. 3 Fusion results, the original image corrupted by the salt and pepper noise and partial blurring. Top: input image X (the original image with added noise and a blurred segment, left), input image Y (the original image with added noise and another blurred segment, middle), fused image F using averaging (right). Bottom: fused image F using ratio pyramid decomposition (left), fused image F using the PCA decomposition (middle), fused image F using DWT domain fusion (right)