New quality measures for image fusion

Gemma Piella
Polytechnical University of Catalonia (UPC)
Jordi Girona 1–3
08034 Barcelona
Spain
piella@gps.tsc.upc.es

Abstract – *We present a new approach for assessing quality in image fusion. The interest of our measures lies in the fact that they do not require a ground-truth or reference image and can be easily computed. We perform simulations which show that our measures are compliant with subjective evaluations and can therefore be used to compare different image fusion methods or to find the best parameters for a given fusion algorithm.*

Keywords: fusion performance, non-reference quality measures, objective quality measures.

1 Introduction

The widespread use of image fusion methods, in military applications, in surveillance, in medical diagnostics, etc., has led to an increasing need for pertinent performance or quality assessment tools in order to compare the results obtained with different algorithms or to obtain an optimal setting of parameters for a given fusion algorithm.

In most cases, image fusion is only a preparatory step to some specific task such as human monitoring, and thus the performance of the fusion algorithm has to be measured in terms of improvement of the subsequent tasks. For example, in classification tasks, a common evaluation measure is the percentage of correct classifications. This requires that the 'true' correct classifications are known. In experimental setups, however, the availability of a ground-truth is not guaranteed.

In this paper, we focus on general performance measures which can be computed independently of the subsequent task. More precisely, we are interested in measures that express the successfulness of an image fusion technique by the extent that it creates a composite image that retains salient information from the source images while minimizing the number of artifacts or the amount of distortion that could interfere with interpretation.

In many applications, the end user of the fusion result is a human. Thus, the human perception of the composite image is of paramount importance and therefore, fusion results are mostly evaluated by subjective criteria [1, 2]. Objective performance assessment is a difficult issue due to the variety of different application requirements and the lack of a clearly defined ground-truth. Indeed, various fusion algorithms presented in the literature (see [3] for an overview) have been evaluated by constructing some kind of ideal composite image and using it as a reference for comparing with the experimental fused results [4, 5]. Mean squared error (MSE) based metrics are widely used for these comparisons.

A restricted number of objective fusion performance measures have been proposed where the knowledge of ground-truth is not assumed. Xydeas and Petrović [6] propose a metric that evaluates the relative amount of edge information that is transferred from the input images to the composite image. In [7], mutual information is employed for evaluating fusion performance.

This paper discusses a novel objective non-reference quality assessment algorithm for image fusion that utilizes local measurements to estimate how well the salient information contained within the sources is represented by the composite image. Our quality measures are based on an image quality index proposed by Wang and Bovik in [8].

2 The image quality index of Wang and Bovik

We start by presenting the image quality index that was introduced by Wang and Bovik in [8]. Given two images a and b of size $M \times N$, let $\bar{a}$ denote the mean of a, let σ_a^2 and σ_{ab} be the variance of a and covariance of a, b, respectively, i.e.,

$$\begin{aligned}\sigma_a^2 &= \frac{1}{MN-1}\sum_{m=1}^{M}\sum_{n=1}^{N}\big(a(m,n)-\bar{a}\big)^2 \\ \sigma_{ab} &= \frac{1}{MN-1}\sum_{m=1}^{M}\sum_{n=1}^{N}\big(a(m,n)-\bar{a}\big)\big(b(m,n)-\bar{b}\big)\,.\end{aligned}$$

Define

$$Q_0 = \frac{4\sigma_{ab}\bar{a}\bar{b}}{(\bar{a}^2+\bar{b}^2)(\sigma_a^2+\sigma_b^2)}\,, \tag{1}$$

which can be decomposed as

$$Q_0 = \frac{\sigma_{ab}}{\sigma_a\sigma_b}\cdot\frac{2\bar{a}\bar{b}}{\bar{a}^2+\bar{b}^2}\cdot\frac{2\sigma_a\sigma_b}{\sigma_a^2+\sigma_b^2}\,. \tag{2}$$

Wang and Bovik refer to Q_0 as an *image quality index* and use it to quantify the structural distortion between images a and b. In fact, the value $Q_0 = Q_0(a,b)$ is a measure for the

similarity of images a and b and takes values between -1 and 1. Note that the first component in Eq. (2) is the correlation coefficient between a and b. The second component corresponds to a kind of average luminance distortion and has a dynamic range of $[0, 1]$ (assuming nonnegative mean values). The third factor in Eq. (2) measures a contrast distortion and its range is also $[0, 1]$. The maximum value $Q_0 = 1$ is achieved when a and b are identical.

Since image signals are generally non-stationary, it is appropriate to measure the number Q_0 over local regions and then combine the different results into a single measure. In [8] the authors propose to use a sliding window approach: starting from the top-left corner of the two images a, b, a sliding window of fixed size moves pixel by pixel over the entire image until the bottom-right corner is reached. For each window w, the local quality index $Q_0(a, b \mid w)$ is computed for the values $a(m, n)$ and $b(m, n)$ where pixels (m, n) lie in the sliding window w. Finally, the overall image quality index Q_0 is computed by averaging all local quality indices:

$$Q_0(a, b) = \frac{1}{|W|} \sum_{w \in W} Q_0(a, b \mid w)\,, \tag{3}$$

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

Wang and Bovik [8] have compared (under several types of distortions) their quality index with existing image measures such as the MSE as well as with subjective evaluations. Their main conclusion was that their new index outperforms the MSE, and they believe this to be due to the index's ability of measuring structural distortions, in contrast to the MSE which is highly sensitive to the l^2 energy of errors.

3 A new fusion quality measure

We use the Wang-Bovik image quality index Q_0 in Eq. (3) to define a quality measure $Q(a, b, f)$ for image fusion. Here a, b are two input images and f is the composite image resulting from the fusion of a and b. The measure $Q(a, b, f)$ should express the 'quality' of the composite image given the inputs a, b.

We denote by $s(a|w)$ some saliency of image a in window w. It should reflect the local relevance of image a within the window w, and it may depend on, e.g. contrast, variance, or entropy. Given the local saliencies $s(a|w)$ and $s(b|w)$ of the two input images a and b, we compute a local weight $\lambda_a(w)$ between 0 and 1 indicating the relative importance of image a compared to image b: the larger $\lambda_a(w)$, the more weight is given to image a. A typical choice for $\lambda_a(w)$ is

$$\lambda_a(w) = \frac{s(a|w)}{s(a|w) + s(b|w)}\,. \tag{4}$$

In a similar fashion we compute $\lambda_b(w)$. Note that in this case $\lambda_b(w) = 1 - \lambda_a(w)$. Now we define the fusion quality measure $Q(a, b, f)$ as

$$Q(a, b, f) = \frac{1}{|W|} \sum_{w \in W} \Big(\lambda_a(w) Q_0(a, f|w) + \lambda_b(w) Q_0(b, f|w) \Big). \tag{5}$$

Thus, in regions where image a has a large saliency compared to b, the quality measure $Q(a, b, f)$ is mainly determined by the 'similarity' of f and input image a. On the other hand, in regions where the saliency of b is much larger than that of a, the measure $Q(a, b, f)$ is mostly determined by the 'similarity' of f and input image b.

At this point, our model has produced a quality measure which gives an indication of how much of the salient information contained in each of the input images has been transferred into the composite image. However, the different quality measures obtained within each window have been treated equally. This is in contrast with the human visual system which is known to give higher importance to visually salient regions in an image. We now define another variant of the fusion quality measure by giving more weight to those windows where the saliency of the input images is higher. These correspond to areas which are likely to be perceptually important parts of the underlying scene. Therefore the quality of the composite image in those areas is of more importance when determining the overall quality. The overall saliency of a window is defined as $C(w) = \max\big(s(a|w), s(b|w)\big)$. The *weighted fusion quality measure* is then defined as

$$Q_W(a, b, f) = \sum_{w \in W} c(w) \Big(\lambda_a(w) Q_0(a, f|w) + \lambda_b(w) Q_0(b, f|w) \Big), \tag{6}$$

where $c(w) = C(w) / \big(\sum_{w' \in W} C(w') \big)$. There are various other ways to compute the weights $c(w)$ (for example, we could define $C(w) = s(a|w) + s(b|w)$), but we have found that the choice made here is a good indicator of important areas in the input images.

We introduce one final modification of the fusion quality measures that takes into account some aspect of the human visual system, namely the importance of edge information. Note that we can evaluate Q_W in Eq. (6) using 'edge images' (e.g. the norm of the gradient) instead of the original grayscale images a, b and f. Let us denote the edge image corresponding with a by a'. Now we combine $Q_W(a, b, f)$ and $Q_W(a', b', f')$ into a so-called *edge-dependent fusion quality index* by

$$Q_E(a, b, f) = Q_W(a, b, f)^{1-\alpha} \cdot Q_W(a', b', f')^{\alpha}\,, \tag{7}$$

where the parameter $\alpha \in [0, 1]$ expresses the contribution of the edge images compared to the original images: the closer α is to 1, the more important is the edge image.

Note that the three proposed measures have a dynamic range of $[-1, 1]$. The closer the value to 1, the higher the quality of the composite image.

4 Experimental results

In this section we use the proposed fusion quality measures in Eqs. (5–7) to evaluate different multiresolution (MR) image fusion schemes. The MR-based image fusion approach consists of performing an MR transform on each input image and, following some specific rules, combining them into a composite MR representation. The composite image is obtained by applying the inverse transform on this composite MR representation [3].

In this paper we only use the Laplacian pyramid, the ratio pyramid and the discrete wavelet transform (DWT), and in all cases we perform a 3-level decomposition. We combine the coefficients of the MR decomposition of each input by selecting at each position the coefficient with a maximum absolute value, except for the approximation coefficients from the lowest resolution where we take the average. For comparison, we also use the simple fusion method of averaging the input images.

In the computation of the measures defined in last section, we take $\lambda_a(w)$ as in Eq. (4), with $s(a|w)$, $s(b|w)$ being the variance (or the average in the edge images) of images a and b, respectively, within the window w of size 8×8. In all displayed images, we perform a histogram stretching and scale the gray values of the pixels between 0 (black) and 255 (white).

First, we take as input images the complementary pair shown in the top row of Fig. 1. They have been created by blurring the original 'Einstein' image of size 256×256 with a disk of diameter of 11 pixels. The images are complementary in the sense that the blurring occurs at the left half and the right half, respectively. In the second row we display their total weights used to compute Q_W in Eq. (6). More specifically, each pixel (m, n) in the left image contains the value $c(w)\lambda_a(w)$ with w being the window whose top-left corner corresponds to (m, n). Similarly, the right image displays $c(w)\lambda_b(w)$ for every $w \in W$. The composite images obtained by the Laplacian pyramid, the ratio pyramid, the DWT and the average are depicted in the third and fourth row, from left to right. Table 1 compares the quality of these composite images using our proposed quality measures. The first row corresponds to the fusion quality measure Q defined in Eq. (5), the second row to the weighted fusion quality measure Q_W in Eq. (6) and the third row to the edge-dependent fusion quality measure Q_E in Eq. (7) with $\alpha = 1/2$. For comparison, we also compute the Q_0 and the root mean squared error (RMSE) between the original 'Einstein' image and each of the composite images. Note that in 'real' fusion scenarios we do not have access to the original image. The resulting errors are shown in the last row of Table 1.

Table 1: Comparison between different quality measures for the composite images in Fig. 1.

measure	Laplacian	Ratio	DWT	Average
Q	**0.901**	0.790	0.892	0.864
Q_W	**0.929**	0.830	0.924	0.901
Q_E	**0.845**	0.668	0.839	0.745
Q_0	**0.987**	0.818	0.976	0.875
RMSE	**1.526**	14.28	2.384	7.862

Fig. 1 shows that the Laplacian and DWT methods are comparable and that they outperform the other two schemes. Note, for instance, the blurring (e.g., in the eyes) and the loss of texture (e.g., the stripes in his clothes) of the composite images obtained by the ratio pyramid and averaging. Furthermore, in the ratio-pyramid composite image, some details of the man's face have been cleared out, and in the average composite image, the loss of contrast is evident. These subjective visual comparisons are corroborated by the results in Table 1. Note that the Laplacian method has higher quality measures than the DWT. This is most likely due to the fact that the former method is better able to preserve edges and reduce the ringing artifacts around them.

Consider now the input images in the top row of Fig. 2. They correspond to a computer tomography (CT) image and a magnetic resonance image (MRI). We repeat the same computations as described above. The results are shown in Fig. 2 and Table 2. Here, however, as we do not have a reference image to compare with, we cannot compute the RMSE nor the Q_0. Instead, we use a measure based on mutual information. More precisely, the results in the last row of Table 2 have been obtained by adding the mutual information between the composite image and each of the inputs, such as in [7], and dividing it by the sum of the entropies of the inputs, i.e.,

$$MI(a, b, f) = \frac{I(a, f) + I(b, f)}{H(a) + H(b)}$$

where $I(a, f)$ is the mutual information between a and f, and $H(a)$ is the entropy of a. In this way, we normalize the measure to the range $[0, 1]$.

Table 2: Comparison between different quality measures for the composite images in Fig. 2.

measure	Laplacian	Ratio	DWT	Average
Q	0.652	0.595	**0.655**	0.629
Q_W	**0.773**	0.651	0.706	0.626
Q_E	**0.797**	0.601	0.733	0.589
MI	0.337	0.221	0.409	**0.691**

In Fig. 2, we can see that again the Laplacian and DWT methods clearly outperform the other two methods. For both of them, many details (specially the brain tissue in the magnetic resonance image) have been lost. Moreover, due to the high contrast in the input images, the ratio pyramid blows up the dynamic range for some pixels, which makes it necessary to clip them in order to be able to 'visualize' the image. Again, the subjective visual analysis is consistent with the new quality measures, as shown in Table 2. In both experiments, the edge-dependent fusion quality index gives a stronger separation between the good results (Laplacian and DWT) and the bad results (ratio and average). Note that the last row, where mutual information has been used, gives the best ranking to the average fusion method. However, mutual information has been shown to be a good indicator of the quality of MR composite images [7] (as long as the average is not taken in all levels for the construction of the composite MR decomposition).

It is interesting to see to what extent edge and illumination influence fusion quality. For that purpose, we have computed the quality measure Q_E varying the parameter α (see Eq. (7)). We have found that for out-of-focus images, like in the 'Einstein' case in Fig. 1, the influence of illumination seems more important than the information provided

by the edges, hence increasing the parameter α would, in general, decrease the value of Q_E. Despite this fact, the rating order between the different fusion methods is preserved. On the other hand, for complementary images where few redundant information is present, as in the 'Skull' images in Fig. 2, the higher the contribution of the edge images, the higher the value Q_E when the fused image has good 'visual' quality . This suggests that for this latter case, the edge information plays an important role. Table 3 shows the values of Q_E obtained with the fused images in Fig. 2 when varying α.

Table 3: Quality measure Q_E when varying the edge fused images contribution.

Q_E	Laplacian	Ratio	DWT	Average
$\alpha = 0.2$	**0.783**	0.630	0.716	0.611
$\alpha = 0.4$	**0.796**	0.610	0.727	0.596
$\alpha = 0.6$	**0.802**	0.591	0.738	0.581
$\alpha = 0.8$	**0.812**	0.573	0.749	0.567
$\alpha = 1$	**0.822**	0.555	0.760	0.553

5 Conclusions

In this paper we have discussed some new objective quality measures for image fusion which do not require a reference image and correlate well with subjective criteria as well as with other existing performance measures. Our measures are easy to calculate and applicable to various input modalities (and hence to different fusion applications). In particular, our measures give good results on variable quality input images since they take into account the locations as well as the magnitude of the distortions.

Further research is necessary to study the influence of the different parameters of the measures (e.g., size of the window, choice of saliency and weights, etc.), and how to select them in order to optimize the quality measures.

There are several areas in which our quality measures can be extended. We currently consider grayscale images, so inclusion of colour is an obvious extension. Other visual mechanisms of our human visual system may also be taken into account. One such mechanism is multiresolution. Since the sensivity of the human visual system varies over spatial frequencies, it seems natural to compute the quality measures with respect to the scales of the objects that appear in the image. Another possible extension is motivated by our work in region-based fusion [3, 9]. Rather than calculating the quality measure in fixed windows, one might choose to segment the sources first and compute the measure region by region.

In addition, we plan to include some information-theoretic measures such as mutual information and entropy to better estimate the information content of the composite image. We also plan to study how our objective measures can be used to guide a fusion algorithm and improve the fusion performance.

Acknowledgment

The author would like to thank Henk Heijmans for his valuable feedback and interesting discussions on the subject.

References

[1] D. Ryan and R. Tinkler. Night pilotage assessment of image fusion. In *Helmet and Head-Mounted Displays and Symbology Design Requirements II, Proc. SPIE*, volume 2465, pages 50–67, Orlando, Florida, April 18-19 1995.

[2] A. Toet and E. M. Franken. Perceptual evaluation of different image fusion schemes. *Displays*, 24(1):25–37, February 2003.

[3] G. Piella. A general framework for multiresolution image fusion: from pixels to regions. *Information Fusion*, 9:259–280, December 2003.

[4] H. Li, B. S. Manjunath, and S. K. Mitra. Multisensor image fusion using the wavelet transform. *Graphical Models and Image Processing*, 57(3):235–245, May 1995.

[5] O. Rockinger. Image sequence fusion using a shift invariant wavelet transform. In *Proc. IEEE Int. Conf. Image Processing*, volume 13, pages 288–291, 1997.

[6] C. Xydeas and V. Petrović. Objective pixel-level image fusion performance measure. In *Sensor Fusion: Architectures, Algorithms and Applications IV, Proc. SPIE*, volume 4051, pages 88–99, Orlando, Florida, April 24-28 2000.

[7] G. H. Qu, D. L. Zhang, and P. F. Yan. Information measure for performance of image fusion. *Electronic Letters*, 38(7):313–315, 2002.

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

[9] G. Piella. A region-based multiresolution image fusion algorithm. In *Proc. Fifth Int. Conf. Information Fusion*, pages 1557–1564, Annapolis, Maryland, July 8–11 2002. Int. Soc. Information Fusion, Sunnyvale, CA, 2002.

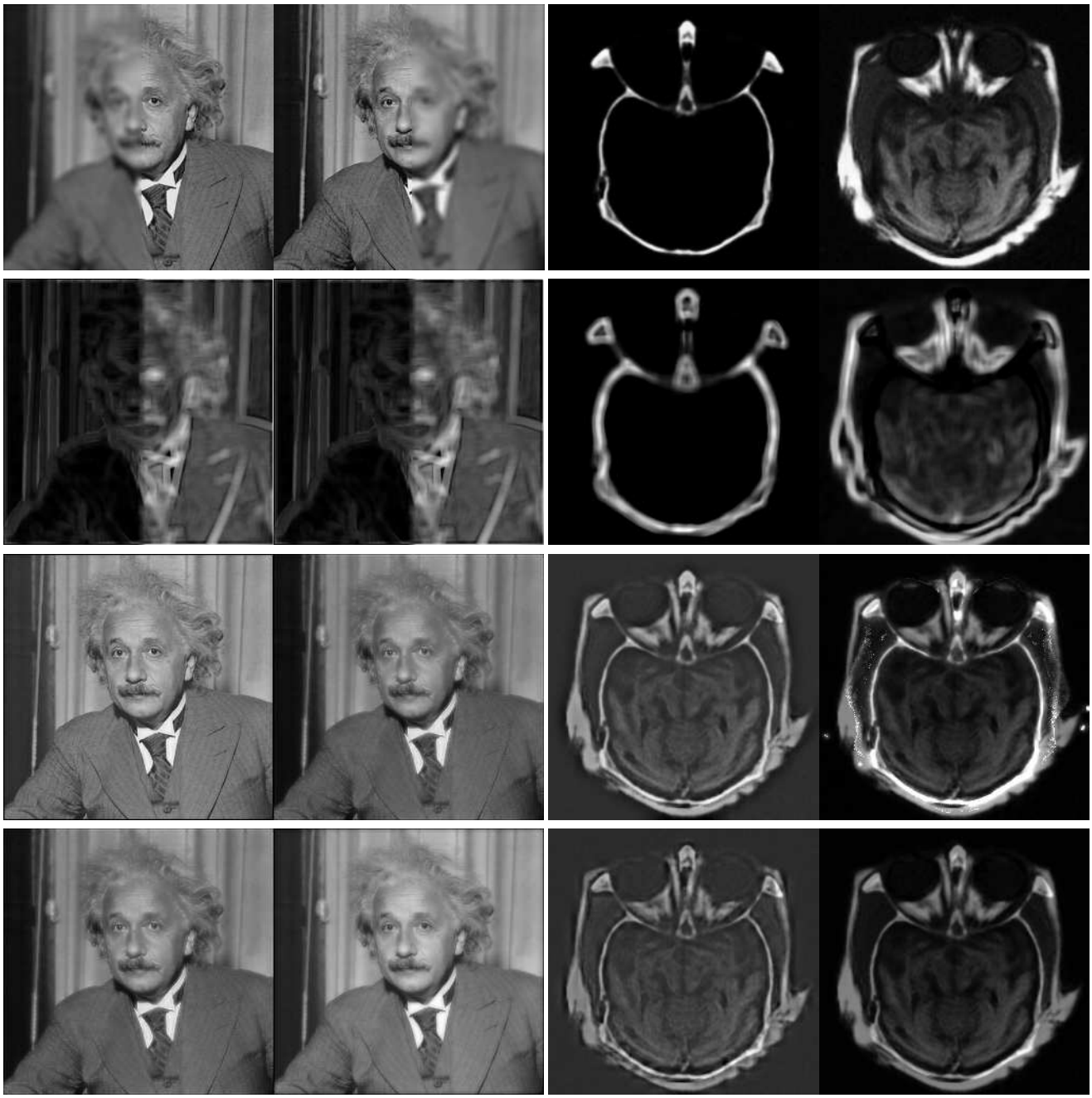

Fig. 1: Experiment 1. Top: input images a (left) and b (right); second row: total weights $c \cdot \lambda_a$ (left) and $c \cdot \lambda_b$ (right); third row: composite images with a Laplacian (left) and a ratio (right) pyramid decompositions; bottom: composite images with a DWT (left) decomposition and averaging (right).

Fig. 2: Experiment 2. Top: input images a (CT image, left) and b (MRI image, right); second row: total weights $c \cdot \lambda_a$ (left) and $c \cdot \lambda_b$ (right); third row: composite images with a Laplacian (left) and a ratio (right) pyramid decompositions; bottom: composite images with a DWT (left) decomposition and averaging (right).