Abstract – The considerable number of image fusion algorithms available today vary widely in terms of fusion performance and robust fusion assessment tools have become a target of considerable research. Based on a variety of localised or global evaluations of image statistics and structure between the inputs and the fused image, available objective fusion evaluation metrics use a number of different information representation and information loss models. This paper explores the definition of an optimal information representation for the evaluation of multisensor image fusion and how it can be defined based on the actual application of fused information. Extensive evaluations of a considerable data set of subjectively annotated fused images with a variety of information representation approaches implemented using three different global fusion evaluation frameworks are presented. The results show that if used with correct information representation models global, statistical approaches can yield significantly better fusion evaluation performance than existing methods.

Keywords: image fusion evaluation, information representation, information loss modelling, objective fusion metrics.

1 Introduction

Multisensor imaging arrays are increasingly becoming commonplace in a growing range of both specific and general imaging applications. However, in order to fully exploit additional information provided by different sensors, considerable processing effort is required. Furthermore, displaying multiple image modalities to a human operator simultaneously leads to confusion and overload, while integrating information across a group of users is almost impossible [1].

Signal-level image fusion has established itself as a useful tool in reducing this information overload for a range of multisensor imaging applications from military avionics to medical imaging. It can provide highly efficient representations of multimodal information to expectant observers as well as highly sensitive, autonomous image analysis systems by reducing the physical amount of multisensor (image) data while preserving its information content value [2-5].

Whereas obtaining a fused image is relatively straightforward, e.g. simply average the inputs, assessing the performance of fusion algorithms, particularly those intended to produce a visual display is much harder in practice. The most reliable and direct method of fusion for display evaluation are subjective trials in which audiences of intended users evaluate fused images under tightly controlled conditions either by comparing them to each other or by performing specific visually oriented tasks [1,6,7]. Subjects’ responses or task performance are logged and need to be processed further to obtain a meaningful performance evaluation making the whole process expensive in terms of time, effort and equipment required.

Objective fusion evaluation metrics in contrast, require no display equipment or audience organisation and their advantage is obvious in terms of evaluation efficiency. Implementation of such algorithms in computer code and simulation experiments reduces assessment time from days or weeks to a few seconds or minutes. A significant advantage of this is that objective metrics can be used to optimise fusion system parameters to a particular input set or type of data expected in an application, in a way that is impossible to achieve using human subjects and complex visual testing procedures. However, a lack of reliable methods to validate their performance has long been a drawback in the design of objective fusion evaluation metrics. But as recent studies show [7], wider exploration into detailed design of objective fusion evaluation metrics is possible using data from real subjective trials.

This paper addresses the issue of defining the optimal information representation for robust evaluation of image fusion performance. The idea is to provide a structured analysis of the relative merits of three global approaches to image information fusion evaluation: the global, local and pixel based and determine what properties of visual information are important for the evaluation of fusion for display systems. The next section examines currently available, objective fusion metrics in more detail and focuses on the various approaches to image fusion evaluation. Section three gives an overview of the three broad information evaluation frameworks while section four defines the various processing methods that can be used to represent the visual information more robustly. Subjective relevance assessment of the presented fusion evaluation frameworks and information representation
models is provided in section five. The paper is concluded in section six.

2 Background

Although it has been steadily gaining in research popularity only a few objective image fusion performance metrics have been published [3,4,8-12]. In particular, most of the metrics published have been relatively application dependent and focused on particular aspects of various fusion scenarios. One of the earliest metrics proposed in [8] examined target signature consistency as an evaluation criterion for detection/recognition applications that however depends on a subjective quality element. A popular approach [3,4] has been to compare a fused image obtained from a fusion algorithm to an "ideally" fused image and estimate performance based on their difference. This is of limited application as it depends on the existence of a ground truth fused image [3,4] which is only ever available in very specific applications such as fusion of multifocus images (where it can be obtained manually).

A more general metric, the $Q_{AF}$, based on the evaluation of the similarities between gradients in the inputs and the fused image using a biologically inspired non-linear response was proposed in [9,10]. Another metric with general appeal is based on the concept of mutual information, known to be well suited to multimodal data assessment [11]. Local intensity and gradient statistics of input and fused images, defined on non-overlapping image tiles are compared as a basis for fusion evaluation in a series of metrics proposed in [12].

More recently, the concept of visual differences, used in image quality evaluation has also been successfully applied to fusion evaluation. In [13] it is shown that probabilities of visual differences between the inputs and the fused image can add further robustness to existing fusion evaluation algorithms, while in [14] contrast sensitivity functions are used to determine the relative quality of local image regions based on the visible differences between the inputs and the fused images.

Finally, interest has also been stirred in theoretic evaluation of fusion performance aimed at setting the theoretic limits for fusion performance and putting experimental evaluation in proper context [15].

3 Fusion evaluation frameworks

All of the image fusion metrics with global appeal presented so far [9-12] exhibit a broadly similar evaluation approach outlined by the structure on Figure 1. In each case a minimum of two input images and a fused image are initially processed to extract some properties considered to be a good representation of the visual information contained in each image. The properties, usually expressed in the form of image structure or intensity parameters are compared between each of the inputs and the fused image using a pre-defined similarity model. This model evaluates the relative merit of the manner in which the information from a particular input image is represented in the fused image. This comparison produces a numerical result that quantifies the success of fusion for the considered data.

![Figure 1: Evaluation approach of the proposed image fusion evaluation frameworks (metrics)](image)

Finally, approaches that measure image similarities on only sections of the scene (image) at a time require some form of aggregation that summarises local scores across the imaged scene into a final fusion performance score.

In the context of this paper we seek an optimal image fusion evaluation framework based on the broad structure on Figure 1 using three basic global approaches: global, localised and pixel-based.

3.1 Global Evaluation - MI

Image fusion evaluation based on the concept of mutual information is the best known example of global statistical evaluation. Mutual information metrics consider the statistics of the entire scene at once to produce one global estimate of fusion success [11]. The most important effect of this is that all parts of the scene are considered to have the same priority.

Mutual information metrics evaluate fusion by calculating the mutual information between each input image and the fused, $I_{AF}$ and $I_{BF}$ from marginal distributions of each image, $P_A(a)$ and the joint distributions between each input and the fused image $P_{AF}(a,f)$:

$$I_{AF} = \sum_{a,f} P_{AF}(a,f) \frac{P_{AF}(a,f)}{P_A(a)p_f(f)}$$ (1)
where \( a \) and \( f \) are representations of image information. Total fusion score is the sum of the mutual information between the inputs and the fused image [11]:

\[
M_F^{AB} = I_{AF} + I_{BF}
\]  

(2)

Fusion evaluation provided by the mutual information approach is based on the co-occurrence of particular image properties across the scene and local structure is not considered. In fact spatial image structure is completely ignored.

One important practical aspect of probability density based approaches, such as this, is the resolution of the measured distributions, or the problem of an optimal number of histogram bins one needs to use to achieve a reliable result. In [11] a full range of 256 bins is proposed for distributions of image intensities (8 bit). However, this might not be optimal as under normal conditions people generally only distinguish about 64 shades of gray and a denser evaluation might unnecessarily confound the evaluation. We test the validity of this assumption in section five by varying the resolution of the measured distributions.

In its basic form [11], image information is raw intensities and mutual information is evaluated from image intensity distributions. In the context of Figure 1 this means that the first box is an identity filter while the other two stages are combined in a single step described by equation (1). MI evaluation framework however, can also consider other image properties as we will see in the next section.

3.2 Localised Area Evaluation

Image fusion metrics based on a localised evaluation approach [12, 14] assume that it is the local image structure that carries the information. They extract image properties and measure similarities between corresponding neighbourhoods of pixels in the inputs and the fused image. Methods such as [12] represent image information through local intensity statistics defined over a number of distinct windows of pixels tiled over the entire image, Figure 2. The similarity is evaluated between each pair of matched neighbourhoods in an input and the fused image using local signal statistics [12]:

\[
Q_0(a, f) = \frac{4\sigma_a \bar{a} \bar{f}}{(\bar{a}^2 + \bar{f}^2)(\sigma_a^2 + \sigma_f^2)}
\]  

(3)

These local scores for the tiles are aggregated across the scene and both inputs using perceptual weighting defined by local image variance [12]:

\[
Q_0^{ABF} = \sum_{n,F} (\lambda(w)Q_0(a, f|w) + (1 - \lambda(w))Q_0(b, f|w))
\]  

(4)

Stated advantage of this method is that it considers local image structures whose integrity in the fusion process is vital for correct representation of the inputs. Like MI, the approach summarised in [12] can also be applied to visual information representations other than raw intensities and is adopted here to test various information representation techniques as applied to localised fusion evaluation.

3.3 Pixelwise Evaluation

Pixelwise evaluation defines a comparison between the inputs and the fused image at every pixel across the scene. The \( Q_{ABF}^{AF} \) framework proposed in [9,10] makes no explicit attempt to measure or extract spatial image structure but rather relies on aggregation of a large number of individual (pixel based) measurements for sufficient structural representation. Similarity evaluation is based on direct or relative differences (changes) in representation parameters between the input and the fused image. This relative change is then mapped using a non linear response model to provide a final information representation (similarity) estimate for that image location:

\[
Q_{n,m}^{AF} = \Gamma(1 + e^{-k(x_{n,m}^{AF} - \sigma)})^{-1}
\]  

(5)

One such non linear response is shown on Figure 3 where \( \Delta \) is a measured change in image property, e.g. gradient intensity, at any one pixel between an input and the fused image and \( Q \) is the resulting similarity. Pixel similarity estimates obtained in this manner are aggregated across the entire scene in a weighted summation where each pixel is assigned some value proportional to its relative importance (analogous to equation (4)).

In [9,10] gradient parameters are used to represent visual information in the images and change parameters were derived as a relative change in gradient strength and absolute change in gradient orientation at every pixel. Other information representations however are also suitable for this evaluation approach and information loss...
estimation. A number were tested by recording the relative change in their values between the inputs and the fused image, processing them with the non-linear response and aggregating them over the entire scene.

4 Information representation

Information representation step in Figure 1 is effectively a mapping of the camera measured illumination by a particular representation model into a domain more suitable for image information analysis. The space of such mapping models and representations is infinite, however most of the models do not relate to the manner in which people or indeed image analysis systems process information. A number of different representations have been used in fusion evaluation so far [3,4,9-14]. The goal here is to test these as well as a number of other more accessible and significant ones, but by no means an exhaustive list, within the context of the three evaluation frameworks presented in the previous section. For the sake of efficiency, we limit our search to representations with no more than 2 dimensions which includes most direct and gradient based representations. Table 1 summarises the methods tested (I stands for image intensity while $s_x$ and $s_y$ are Sobel gradients).

<table>
<thead>
<tr>
<th>Structure Representation</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>intensity</td>
<td>$I$</td>
</tr>
<tr>
<td>mean normalised intensity</td>
<td>$I - E[I]$</td>
</tr>
<tr>
<td>unit normed intensity</td>
<td>$(I - E[I])/\sigma^2$</td>
</tr>
<tr>
<td>gradient orientation</td>
<td>$arctan(s_x/s_y)$</td>
</tr>
<tr>
<td>gradient magnitude, $s_g$</td>
<td>$s_x^2 + s_y^2$</td>
</tr>
<tr>
<td>gradient contrast</td>
<td>$s_x E[s_y]/w$</td>
</tr>
<tr>
<td>Sobel gradients</td>
<td>$(s_x, s_y)$</td>
</tr>
<tr>
<td>normalised gradients</td>
<td>$\left(\frac{s_x}{s_x^2 + s_y^2}, \frac{s_y}{s_x^2 + s_y^2}\right)$</td>
</tr>
<tr>
<td>square mapped gradients</td>
<td>$\left(\frac{s_x^2 - s_y^2}{s_x^2 + s_y^2}, \frac{2s_x s_y}{s_x^2 + s_y^2}\right)$</td>
</tr>
</tbody>
</table>

Direct information representations such as raw image intensities have been the most widely used representation in fusion evaluation [3,4,11,12] but they tend to suffer from noise and widely varying local statistics. Some normalised intensity representations: mean normalised (brought to zero mean) and unit normed (zero mean and unit variance) intensity overcome these problems and are also included.

Another important approach to representation of visual information contained within an image is based on changes in image intensity rather than intensities directly. It is the changes in intensity across the images that give rise to edges and patterns that we perceive as image structure. Basic properties of image signal change are amplitude and orientation. Local contrast, evaluated at every pixel as the ratio of image intensity at that location and mean background illumination measured over a local (5x5) neighbourhood of pixels $w$. Another way to measure local amplitude of signal change is through gradient magnitude, evaluated from Sobel horizontal and vertical gradient responses $s_x$ and $s_y$. Change, or gradient orientation captures the shape component of image structure and is also measured from Sobel gradient responses. Gradient contrast, as a higher order statistic evaluated as a ratio of gradient magnitude at each pixel and average gradient magnitude in a 5x5 neighbourhood around the pixel, emphasises highly non linear image areas.

Gradients are two dimensional representations of image structure that are known to have a direct bearing on how observers perceive images and have been used in fusion evaluation in the past [9,10,12]. In order to determine what aspect of gradient information is important in evaluating fused images, several gradient representations were tested including simple Sobel gradient responses, normalised gradients (gradient vectors normalised to unit length) and square mapped gradients (normalised gradients which exhibit more robustness to noise). Finally a further two dimensional representation using gradient strength and orientation derived from Sobel gradient responses, originally proposed in [9,10] was also tested.

A subset of tested representations is shown on Figure 4 for an example image from the multisensor data set used in this investigation, 4a. Although there is no visible noise distortion, it is evident that raw image intensities in 4a
exhibit widely differing local statistics across the scene. Gradient orientation in 4b appears much more random due to scaling but is in fact limited to a range \([-\pi/2, \pi/2]\) from horizontal (back) through vertical (gray) to inverted horizontal (white). Local contrast representation in Figure 4c has a limited range and is only significant close to sharp edges. Gradient magnitude in 4d shows a much better behaved signal with evenly distributed responses that clearly localises all areas of saliency. Finally, Sobel x and y gradients in 4e and 4f also localise important information well but also separate image features into horizontal and vertical components.

5 Results

Assessing the accuracy of fusion performance metrics is a difficult task that ideally requires a calibrated set of results against which the proposed metric can be compared. One such set is available from subjective trial results reported in [7] that were run on an extensive set of multisensor images fused using a number of different image fusion algorithms. In the trials, observers were shown a multisensor image pair and two fused versions of the pair and asked to decide which if any of the two offered fused images better represents the scene shown in the inputs. Subjects expressed their preference or lack thereof by voting for one or none of the fused images. An example of the images used in the subjective trials is shown in Figure 5. Details of the test procedure and conditions as well as trial results can be found in [7].

![Figure 5: Example input (top) and fused images (bottom)](image)

Given such subjective trial results, a perceptually relevant fusion metric should be able to predict them with reasonable accuracy. They are thus used to evaluate the different information representations within the context of subjective relevance of image fusion evaluation they provide. In order to obtain a quantitative evaluation, subjects’ preference votes for each pair of fused images are aggregated and normalised by the number of subjects to give preference scores \(S_1, S_2\) and \(S_0\) (no preference) respectively, \(S \in [0,1]\). Both fused images are then evaluated using the metric and an objective preference is recorded \(O_p=1, p \in \{1,2\}\) for the image with the higher metric score, or no preference \((O_p=0)\) if the scores are within 1.5% (found to be sufficient given the limited practical range of the metrics) [7]. From subjective \((S_0, S_1, S_2)\) and objective \((O_0, O_1, O_2)\) preference scores for each evaluated fused image pair, two distinct measures of subjective relevance are derived [7]. Correct Ranking measure \(C\) is the proportion of all pairs for which metrics ranked the fused images the same way as the subjects. A value of unity means ideal agreement. Relevance measure \(r\), equation (6), takes into account the relative certainty of subjective ranking. By summing subjective preference scores for the fused image indicated as being the better in each pair by the objective metric. Relevance is larger when the metric correctly ranks pairs in which subjects are more unanimous. Additionally, \(r\) is scaled to \([0,1]\) by the sum of maximum preferences.

\[
r = \frac{\sum_{i=1}^{N} \sum_{k=0}^{2} S_i O_k}{\sum_{i=1}^{N} \max(S_i', S_i', S_i')}
\]

Correct ranking, \(C\), and relevance, \(r\), results for the three evaluation frameworks implemented using representation strategies in Table 1 are given in Tables 2 and 3 with best scores in bold. The results show that subjective relevance of the tested evaluation frameworks greatly depends on the choice of information representation. The approaches perform optimally on different information representations with greatest subjective relevance achieved by the Mutual information evaluation based on gradient magnitude representation. This system achieves a correct ranking rate \(C\) 5% higher to the next best system (pixel based using magnitude+orientation [9,10]). It also shows a significant improvement in relevance \(r\) (from 0.833 to 0.87).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>intensity</td>
<td>72.5 %</td>
<td>67.5 %</td>
<td>30 %</td>
</tr>
<tr>
<td>mean norm. intensity</td>
<td>66.7 %</td>
<td>71.7 %</td>
<td>56.7 %</td>
</tr>
<tr>
<td>unit norm. intensity</td>
<td>65 %</td>
<td>62.5 %</td>
<td>62.5 %</td>
</tr>
<tr>
<td>local contrast</td>
<td>70.8 %</td>
<td>50 %</td>
<td>18.9 %</td>
</tr>
<tr>
<td>gradient orientation</td>
<td>60.8 %</td>
<td>63.3 %</td>
<td>52.2 %</td>
</tr>
<tr>
<td>gradient magnitude</td>
<td>77.5 %</td>
<td>67.5 %</td>
<td>69.2 %</td>
</tr>
<tr>
<td>magnitude &amp; orient.</td>
<td>60.8 %</td>
<td>62.5 %</td>
<td>72.5 %</td>
</tr>
<tr>
<td>gradient contrast</td>
<td>63.3 %</td>
<td>65 %</td>
<td>55 %</td>
</tr>
<tr>
<td>Sobel gradients</td>
<td>70 %</td>
<td>68.3 %</td>
<td>67.5 %</td>
</tr>
<tr>
<td>normalised gradients</td>
<td>60.8 %</td>
<td>63.3 %</td>
<td>33.3 %</td>
</tr>
<tr>
<td>square mapped grads</td>
<td>60.8 %</td>
<td>60.8 %</td>
<td>32.5 %</td>
</tr>
</tbody>
</table>
Table 3: Relevance $r$ results of the tested systems

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>intensity</td>
<td>0.847</td>
<td>0.792</td>
<td>0.414</td>
</tr>
<tr>
<td>mean norm. intensity</td>
<td>0.755</td>
<td><strong>0.829</strong></td>
<td>0.668</td>
</tr>
<tr>
<td>unit norm. intensity</td>
<td>0.732</td>
<td>0.754</td>
<td>0.724</td>
</tr>
<tr>
<td>local contrast</td>
<td>0.837</td>
<td>0.620</td>
<td>0.354</td>
</tr>
<tr>
<td>gradient orientation</td>
<td>0.755</td>
<td>0.773</td>
<td>0.726</td>
</tr>
<tr>
<td>gradient magnitude</td>
<td><strong>0.87</strong></td>
<td>0.776</td>
<td>0.792</td>
</tr>
<tr>
<td>magnitude &amp; orient.</td>
<td>0.753</td>
<td>0.758</td>
<td><strong>0.833</strong></td>
</tr>
<tr>
<td>gradient contrast</td>
<td>0.773</td>
<td>0.774</td>
<td>0.679</td>
</tr>
<tr>
<td>Sobel gradients</td>
<td>0.826</td>
<td>0.803</td>
<td>0.781</td>
</tr>
<tr>
<td>normalised gradients</td>
<td>0.755</td>
<td>0.758</td>
<td>0.442</td>
</tr>
<tr>
<td>square mapped grads</td>
<td>0.761</td>
<td>0.748</td>
<td>0.441</td>
</tr>
</tbody>
</table>

In general, the global mutual information approach seems to offer a more robust performance compared to the other two approaches, for a range of different information representations. These results however are reported for optimal histogram resolution applied to evaluate the marginal and joint distributions of the input and fused signals across the image set. Figure 6 demonstrates performance dependence on this parameter of the MI metric using raw image intensities and gradient magnitude. For intensities a crude distribution evaluation with few bins produces better results while gradient magnitude representation is less sensitive, giving an optimum of around 52 (for a full range of 0 to 255).

Figure 6: Subjective relevance $r$, of the MI metric based on raw intensities (dashed) and gradient strength (solid) against histogram resolution

A similar issue exists with the pixel-based evaluation where the performance depends on the optimal parameters of the information loss non-linearity. Figure 7 shows the results of optimisation of parameters $\kappa$ and $\sigma$, equation (5) against relevance $r$ for the subjective trial data from [7]. It can be seen that performance is much more sensitive to $\sigma$ which determines the positioning of the curve between left and right edges ($\kappa$ determines the slope), Figure 3. The optimal non-linearity, shown on Figure 3, indicates that subjects are sensitive to changes in gradient magnitude between the inputs and the fused image as even small changes in gradient magnitude cause a significant drop in perceived fusion performance.

Figure 7: Subjective relevance $r$ of the pixel based, gradient magnitude metric against non-linearity parameters $\kappa$ and $\sigma$

In terms of information representations gradients and gradient based information representations generally exhibit reliable performance for global, MI as well smaller scale, pixel-based fusion evaluation. Localised evaluation, based on neighbourhood statistics seems to capture this information sufficiently with normalised intensities although it offers the worst optimal performance overall.

For the example images in Figure 5, the subjects chose the fused image on the right (5d) as a better representation of the inputs by 12 votes to 2 with 5 equal preferences. MI metric with gradient magnitude representation scores in line with this result $M_{5c}=0.691$ and $M_{5d}=0.756$. Likewise, pixel based with gradient magnitude and orientation metric $Q_{AB/F}$ scores $Q_{ab/d}=0.484$ and $Q_{ab/d}=0.564$.

6 Conclusions

This paper presents results of an investigation into optimal image information representations for evaluating multisensor image fusion. A wide range of different representation models was considered for the context of three major fusion evaluation frameworks. It was found that mutual information approach based on global statistics of input and fused images that does not take into account local image structure shows optimal agreement with subjective results. When used in conjunction with local gradient magnitude image representation this approach considerably outperforms all other fusion metrics proposed thus far in terms of subjective relevance.

Such results suggest that globalised evaluation approaches are more relevant in general, situation overview fusion applications such as those used in the
subjective trials used in this study. Observers in this context rarely delve into fine image details used by more localised approaches as basis for fusion evaluation. Preservation of local structure could nevertheless become important in task based evaluation where the display of fine structures relevant to the task would drive subjective and hence fusion performance. Further work that would include such task based subjective evaluation should provide more useful insight into this aspect of image fusion performance.

Another remaining issue that needs to be addressed is the number of bins (resolution) of the marginal and joint distributions used to evaluate this metric as performance is shown to be sensitive to this parameter. Further work will investigate how this parameter can be determined a-priori for any given set of multisensor input image data.

Acknowledgements

The work in this paper was funded by UK’s EPSRC and MRC under the MIAS-IRC grant. Dr. Lex Toet and Canadian Defence Establishment Valcartier are gratefully acknowledged for some of the imagery used in the presented work. Data from the subjective trials including both the images and subject responses used in this investigation are available online at [http://www.isbe.man.ac.uk/~vsp/Subjective_Tests.htm](http://www.isbe.man.ac.uk/~vsp/Subjective_Tests.htm).

References