Psychophysical and Metric Assessment of Fused Images

Timothy D. Dixon,* Jan Noyes, Tom Troscianko
Department of Experimental Psychology
University of Bristol, UK

Eduardo Fernández Canga,† Dave Bull,
Nishan Canagarajah
Department of Electrical & Electronic Engineering
University of Bristol, UK

1 Introduction

In this paper we consider the effects of image compression and fusion on image usability. These concepts are introduced in the current section. Section 2 considers the current experimental method, whilst Section 3 presents experiments and results relating to these concepts, including psychophysical results. (Sections 3.1–3.4) and computational metric results (Section 3.5). Section 4 discusses these findings and draws some conclusions based on the results obtained.

* e-mail: Timothy.Dixon@bristol.ac.uk
† e-mail: E.Fernandez-Canga@bristol.ac.uk

1.1 Image Fusion

Multi-sensor image fusion can be used to combine images with differing spectral or spatial characteristics, for example visible and infrared radiation, or high and low spatial resolution, and create an output that contains either all information, or essential information from the inputs. Thus, image fusion has become a major tool in military (e.g. surveillance [Toet et al. 1997]; situational awareness [Stiles and Hofmann 1997]) and civil (e.g. medical imaging [Tomura et al. 2004]; remote sensing [Pohl and Van Genderen 1998]) fields over the past 20 years. The essential aim of image fusion is a reduction in information size, to a point where only the critical information required for interpretation is present. This can then be analysed by either a human or computer end-user, or a combination of the two.

There are now many different ways to fuse images, from simple pixel-based methods such as averaging each pixel value between two images, to more complex region and decision levels [Pohl and Van Genderen 1998], which include pyramid techniques [Toet 1992], wavelet methods [Kingsbury 1998], as well as methods that cross between these levels [Piella 2003]. With such a wide choice of fusion methods, and with image capture devices capable of creating vast amounts of data, an issue storage space for the data generated arises.

1.2 Compression

Multiple input images as well as potential to create many more fused outputs entails a huge amount of soft storage; therefore compression of the images is desirable in order to reduce this strain. Such compression needs to keep the data as accurate as possible whilst providing maximal capacity reduction. The JPEG standard [Wallace 1992] uses a Discrete Cosine Transform (DCT) to create lossy compressions of the original image. It is known to create a ‘blocking’ effect at high compression levels, and whilst research has found the JPEG standard to be outperformed by other compression methods [Ghafoorian and Huang 1995], its ubiquity and performance in applied situations [Rigolin et al. 1996] suggest that it is an appropriate compression scheme.

A newer standard, JPEG2000 [Taubman and Marcellin 2002], performs well at low bit-rates, with less distortion and better subjective quality ratings than previous standards [Santa-Cruz et al. 2002]. Comparisons between the schemes have suggested that JPEG2000 outperforms JPEG in computational and subjective ratings tests [Skodras et al. 2001; Steingrímsson and Simon 2003], although objective performance comparisons using humans are not common in previous research carried out, creating a gap in the current comparisons between the two standards that is appropriate for exploration.

1.3 Image assessment

The assessment of both compressed and fused images has focused largely on finding appropriate computational metrics that will correlate well with subjective quality assessment, although such met
metrics often fail in finding such correlations [Miyahara et al. 1998; Wang et al. 2004]. Additionally, subjective tasks have been shown to not compare well with objective task performance in certain scenarios [Canga et al. 2005]. An alternative approach is to apply a task to each image rating scenario, with the advantage that objective data is drawn from the sample, in terms of accuracy and reaction times. The current study advances work begun by Toet and colleagues [1997; 2003], utilising an alternative experimental paradigm and accounting for the presence of compression in the fusion process as first presented by Canga and colleagues [2005]. This paradigm involved the use of an energy-masked rapid visual presentation of the UN (United Nations) Camp image sequence originally used by Toet et al. [1997], and are captured in both infrared and visible light wavebands. These images are available publicly at www.imagefusion.org. In this sequence, images showed either a soldier (target present) or showed just the surrounding scene (target absent).

We extended the use of psychophysical testing, combined with metric assessment, to compare the JPEG2000 (Experiment 1) and JPEG (Experiment 2) compression schemes. The fusion methods used are simple averaging (AV), contrast pyramid (CP) [Toet 1992], and the dual-tree contrast wavelet transform (DT-CWT) [Kingsbury 1998].

1.4 Metrics

A number of image quality metrics have been proposed including mean square error (MSE), root mean square error (RMSE), peak signal to noise ratio (PSNR), mean absolute error (MAE) and quality index. All of these require a reference image, which is usually the ideal fused image. However, in practice, such an ideal fused image is rarely known. Hence, other fused image metrics such as mutual information (MI) [Guihong et al. 2001], Petrovic and Xydeas metric [Petrovic and Xydeas 2000; Petrovic and Xydeas 2003] and Piella’s Quality Index [Piella 2003] have been recently proposed. These estimate how and what information is transferred from the input images to the fused image.

1.4.1 Mutual Information

The mutual information is a natural measure of the dependence between random variables. Mutual information was first used for image fusion assessment by Guihong et al. [2001] as an alternative to RMSE. It is defined by Kullback-Leibler ‘distance’ between two images, or, in the case of image fusion, by the average of the distance between the input images and the fused image.

1.4.2 Xydeas and Petrovic Metric

Recently, Petrovic and Xydeas [2000; 2003] proposed a metric, which measures the amount of edge information ‘transferred’ from the source image to the fused image to give an estimation of the performance of the fusion algorithm.

This metric uses a Sobel edge operator to calculate the strength and orientation information of each pixel in the input and output images. These measures are then used to estimate the edge strength and orientation preservation values $O^E(n,m)$ and $Q^E(n,m)$, together with maps $w_1(n,m)$ and $w_2(n,m)$, which reflect the perceptual importance of the corresponding edge elements within the input images. These maps are used to weight the estimates of the edge information, which gives the normalised summation of the performance metric ($Q_p^{M/F}$). Note that in this method the visual information is associated with the edge information while the region information is ignored.

$$Q_p^{M/F} = \frac{\sum_{m=1}^{M} \sum_{n=1}^{N} Q^E(n,m)w_1(n,m) + Q^E(n,m)w_2(n,m)}{\sum_{m=1}^{M} \sum_{n=1}^{N} w_1(n,m) + w_2(n,m)}$$

1.4.3 Image Fusion Quality Index

This image fusion quality index (IFQI) [Piella 2003] is based on the structural similarity (SSIM) image quality index recently introduced by Wang and Bovik [2002], which is defined as:

$$Q = \frac{4\sigma_{nqw}^2}{\sigma_n^2 + \sigma_w^2 + (\bar{x}^2 + \bar{y}^2)^2}$$

This SSIM index measures three different aspects: correlation, luminance distortion and contrast distortion.

In order to apply this metric for image fusion evaluation, Piella and Heijmans [2003] introduce salient information to reflect the relative importance of image $A$ compared to image $B$ within the window $w$.

$$Q_x(A,B,F) = \sum_{v \in G} c(v)\lambda(v)\delta(Q(A,F|v) + (1 - \lambda(v))Q(B,F|v))$$

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’ ($A^{'}, B^{'}$ and $F^{'}$) instead of the grey-scale images $A, B$ and $F$. As with the previous metrics, this metric does not require a ground-truth or reference image.

2 Method

Two psychophysical experiments were carried, with Experiment 1 (Section 2.1) using JPEG2000 images in the test sequences as one variable and Experiment 2 (Section 2.2) using JPEG.

2.1 Experiment 1

2.1.1 Design

The first experiment carried out had two independent variables in a repeated measures design. The first variable was image fusion which had three levels: AV, CP, and DT-CWT. The second independent variable was JPEG2000 compression level, which also had three levels: clean (no compression), low (1:27 compression ratio, or 0.3 bpp), and high (1:40 compression ratio, or 0.2 bpp). The dependent variables were hit rate (number of times participants correctly identified the target present), and false alarm rate (the number of times participants thought the target was present when it was not). Trials were blocked and counterbalanced by fusion type, with compression level randomised within each block. In each condition, three different target-present locations were used, either to the left, central, or to the right of the image, as well as three target-absent images. Each image was shown 10 times in each condition, creating a total of 540 images, blocked into groups of 90 images. A backward mask was applied after each test image, in order to reduce ceiling effects by interrupting further processing of the test image [Breitmeyer and Ogmen 2000].
2.1.2 Participants

The participants in this experiment comprised 16 females and 2 males, with a mean age of 22.2 years (age range: 18.8-36.9 years, s.d. = 5.00 years). All participants had normal or corrected-to-normal vision, and none had any prior knowledge of the topic of the experiment.

2.1.3 Apparatus and Stimuli

The experiment was presented on a 19” flat screen CRT (cathode ray tube) monitor, connected to a 2.8GHz Pentium 4 PC (personal computer) with 512 Megabytes RAM, running Superlab Pro v2.0 by Cedrus [1999]. Responses were given using a standard keyboard. Written instructions were displayed at font size 30 in Arial script, with all presentations centrally aligned. Test images were all monochrome, and displayed at full size (360x270 pixels) against a 50% grey background.

2.1.4 Procedure

Participants were asked to sign a consent form, and then shown the video sequence from which the four test images were taken, first with visible and infrared sequences separate side-by-side, and second with the sequences fused. They were given a brief explanation of what image fusion was, and were told that the following experiment would contain images from the sequence that they had just viewed. Instruction were then displayed on the screen, explaining how each experimental trail would run, and allowing participants to have 12 practice trials with feedback.

Each trial in the experiment began with a ‘+’ fixation point being displayed for 750ms in the centre of the screen. The test image was then displayed for 15ms, followed by a 15ms inter-stimulus interval. The energy mask was then displayed for 250ms, after which the screen went blank, and participants responded, pressing ‘C’ if they thought the target was present, and ‘N’ if they thought it was absent. After every 90 trials there was a self-timed rest period.

2.2 Experiment 2

2.2.1 Design

The second experiment replicated the first, except that the independent variable JPEG2000 compression was replaced with JPEG. This variable again had three levels: clean, low compression (1:27 compression ratio, or 0.3bpp), and high compression (1:40 compression ratio, or 0.2bpp). JPEG compression values were controlled and matched to the JPEG2000 levels by comparing the size of the compressed images with the original, and matching these with JPEG2000 compressed images of the same compression size.

2.2.2 Participants

A further 18 participants (16 females, 2 males) took part in this experiment. Their mean age was 20.1 years (range: 18.7-24.8, s.d. = 1.47).

2.2.3 Apparatus and Stimuli

The same apparatus was used as in experiment one, with the only alteration being that the test stimuli were compressed using the JPEG standard in Experiment 2, as detailed in Figure 2.

2.2.4 Procedure

An identical procedure to Experiment 1 was used.

3 Results

The hit and false alarm rate results from both experiments were used to calculate the signal detection parameters $d'$ (sensitivity) and $\beta$ (bias) (e.g. [Wickens 2002]), where $d'$ represents the distance between signal and noise distributions, and $\beta$ indicates the bias to answer either ‘present’ or ‘absent’. The $d'$ score is calculated by converting hit and false alarm rates to z-scores, indicating the position of a score in relation to the number of standard deviations it is above or below the mean. Thus, $d'$ equals the hit ratio z-score minus the false alarm ratio z-score. The $\beta$ value is calculated by
A two-factor repeated measure ANOVA supported the descriptive statistics. For $d'$, there was a significant main effect of fusion type ($F(2, 34) = 11.1, p < 0.001$), but no compression level ($F(2, 34) = 1.19, p = 0.316$), although there was an interaction between the two factors ($F(4, 68) = 2.52, p = 0.049$), as shown in Figure 3.

Post hoc testing using the Bonferroni test indicated that the AV fusion was not significantly lower than CP (1.59 vs. 2.11, $p = 0.070$), although it is approaching significance on a two-tailed test. A significant difference was found between AV and DT-CWT conditions (1.59 vs. 2.63, $p < 0.001$), but not between CP and DT-CWT. This indicates that the CP and DT-CWT conditions had similar performance patterns to one another, both of which kept the signal and noise responses significantly further apart than the averaging condition.

A Tukey’s Honestly Significant Difference (HSD) test was carried out to locate the interaction. This showed that the CP results significantly varied across compression type, differing from AV in the clean (2.13 vs. 1.60, $HSD = 0.454, p = 0.01$) and high (2.32 vs. 1.53, $p = 0.01$) compression, and differing significantly from DT-CWT in the low compression condition (1.88 vs. 2.64, $p = 0.01$). This indicates that whilst compression is having little overall effect, it does seem to be affecting CP and AV fusion, especially at the low compression level.

### 3.1.2 Bias

The bias levels showed similar trends, as participants were more likely to answer ‘no’ than ‘yes’ in AV condition ($\hat{\beta} = 1.18$), less so but still biased with CP (0.827), and minimally biased towards saying ‘yes’ with the DT-CWT ($-0.008$). However, the compression levels again showed little difference (clean = 0.651, low = 0.672, and high = 0.720), indicating a smaller but regular bias towards giving a ‘no’ answer.
is significantly less bias towards answering ‘no’ in the DT-CWT condition than either of the other two conditions.

Tukey’s Honestly Significant Difference tests were carried out to investigate the direction of the interaction. As indicated in Figure 4, there was a significant difference in the high compression condition between CP and DT-CWT fusion types (1.15 vs. −0.193, HSD = 0.570, p = 0.01). This suggests that high compression was particularly affected by the difference between DT-CWT and CP fusion types; with participants much keener to answer ‘no’ in the high compression and CP condition was used than the high compression and DT-CWT condition. Furthermore, this suggests that high compression combined with DT-CWT significantly decreases participants’ bias in response.

3.2 Experiment 2

3.2.1 Sensitivity

![Figure 5: ANOVA of $d'$ - Fusion and JPEG compression](image)

The signal detection analysis of experiment two results for fusion type showed similar sensitivity trends to experiment one, with DT-CWT having the largest $d'$ ($d' = 2.54$), followed by CP (1.74), and then AV (0.940). However, some difference was also found in the compression results: clean = 1.85, low = 1.72, high = 1.65. A repeated measures ANOVA showed a significant main effect of both fusion type ($F(2, 34) = 30.7, p < 0.001$), and compression level ($F(2, 34) = 3.77, p = 0.033$), although no interaction ($F(4, 68) = 0.837, p = 0.506$), as displayed in Figure 5.

Bonferroni testing showed that AV and CP (0.940 vs. 1.74, $p = 0.005$), AV and DT-CWT (0.940 vs. 2.54, $p < 0.001$), as well as CP and DT-CWT (1.74 vs. 2.54, $p = 0.001$) were all significantly different from one another. Both Bonferroni and Tukey’s HSD were too conservative to reveal any significant differences between compression levels. Thus, planned t-tests were applied, revealing one significant difference between clean and high compression ($t(17) = −2.33, p = 0.032$). This suggests that the effect of compression on these results was very small.

3.2.2 Bias

![Figure 6: ANOVA of $\beta$ - Fusion and JPEG compression](image)

The $\beta$ results showed that participants were most biased towards saying ‘no’ with CP fusion ($\beta = 1.20$), less so with AV (0.946), and much less so with DT-CWT (0.147). Compression bias differed somewhat between levels, with clean images having most bias (0.877), then low compressed (0.784), and high compressed images having least bias (0.628). ANOVA testing showed that these trends were significant, with a main effect of fusion ($F(2, 34) = 8.23, p = 0.001$), as well as compression level ($F(2, 34) = 4.32, p = 0.021$), but no interaction ($F(4, 68) = 0.755, p = 0.558$) which can be seen in Figure 6.

Bonferroni testing revealed significant differences between AV and DT-CWT (0.946 vs. 0.147, $p = 0.027$), and between CP and DT-CWT (1.20 vs. 0.147, $p = 0.001$), but not between AV and CP. A significant difference was also found between high compression and clean image: 0.628 vs. 0.877, $p = 0.029$, although no other compression comparisons yielded significant differences. These results indicate that the CP fusion scheme is creating more bias than even the simple averaging when used in conjunction with the JPEG standard. Additionally, there is a stable pattern of bias reduction as compression rate increases, which, whilst not particularly large, is a somewhat surprising finding.

3.3 Comparison of JPEG and JPEG2000 Results

The results for the JPEG and JPEG2000 conditions were compared using a ‘between-subjects’ ANOVA. Whilst this is not strictly appropriate as fusion type remains a ‘within-subjects’ factor in each experiment, it was decided that the ANOVA is robust enough to cater for systematic variations that might occur. In addition, it is not appropriate in this case to consider the ‘compression’ condition, as this condition varies systematically between the two experiments. Instead, the grouping of either JPEG or JPEG2000 can be analysed as a separate condition. The ANOVA results for $d'$ showed significant effects of fusion type ($F(2, 306) = 68.1, p < 0.001$), which was revealed to be significant at all 3 levels by Tukey’s HSD, as well as group ($F(1, 306) = 15.8, p < 0.001$), showing results for JPEG2000 to have greater $d'$ than for JPEG, and an interaction between the two ($F(2, 306) = 3.05, p = 0.049$), as detailed in Figure 7.

The interaction suggests that fusion type is having more of an effect on the JPEG compressed group of results than the JPEG2000 group. The benefit of using the CT-DWT scheme in terms of $d'$ scores for the JPEG group is nearly the same as the JPEG2000 group, suggesting that the CT-DWT scheme is capable of producing images that are much more easily identifiable than the other two schemes, with the AV scheme performing especially poorly when combined with JPEG compression.

The $\beta$ results showed a different pattern, with the only significant main effect being for fusion type ($F(2, 306) = 46.2, p < 0.001$), with DT-CWT significantly lower than CP (0.009 vs. 1.03, $p <
3.4 Discussion of Psychophysical Results

It is clear that different patterns of results are arising when either JPEG or JPEG2000 compression is used. The effect of fusion type appears to remain quite stable, with DT-CWT outperforming the other schemes in all cases, and CP outperforming AV in all but one case. However, when JPEG2000 compression is used, there is no main effect of compression type on participants’ responses (Figures 3 and 4), whereas when JPEG compression is used this effect is present (Figures 5 and 6). This might indicate that whilst controls were in place to ensure that JPEG and JPEG2000 levels were equivalent in terms of the compression applied, JPEG compression still created a noisier image than JPEG2000, most probably due to the characteristic blocking effect seen in JPEG compression.

Another difference in the results relates to the pattern of interactions found, with the JPEG2000 results interacting between compression level and fusion type, whilst the JPEG results did not. This is most obvious in the JPEG2000 results for CP, wherein some interesting effects at low compression levels have affected the $d'$ results (Figure 3), and at high compression levels have affected the $\beta$ results (Figure 4). This might be due to a characteristic of the CP fusion scheme interacting with wavelet compression used by the JPEG2000 standard, and implies that this combination of compression and fusion is not as stable as the DTCWT scheme.

One final interesting finding is that the JPEG results for $\beta$ (Figure 6) show a systematic decrease in bias as the compression levels rise. Whilst this decrease is not particularly large, it is worth noting that participants are not reacting in the same way in this case, perhaps being forced to concentrate more on whether the target was present or not by the distortion caused, and therefore not responding with such bias towards saying ‘no’.

3.5 Metric Results

This section presents the results obtained by computing three different metrics: mutual information, image fusion quality index and Petrovic’s metric. All of them have been computed by using the clean input images together with the image obtained by fusing the compressed images. None of these metrics account for the quality of input images; thus clean images prior to compression are used in these metrics so that the effects of compression on image quality are also considered [Canga et al. 2005].

The results obtained with mutual information are presented in Figure 8. The solid line represents the values obtained with JPEG2000 while the dashed line represents the values obtained with JPEG. It can be observed that, according to this metric and in contradiction to the psycho-visual results, AV fusion method outperforms the other two fusion methods both in JPEG and JPEG2000 compression. JPEG2000 compression seems to have little influence in fusion performance. However, in the case of JPEG compression, a surprising effect on the AV fusion method can be observed, whereas the presence of compression seems to improve the quality of the fusion process.

Figures 9 and 10 show the results of Petrovic’s metric and IFQI measurements respectively. The two of them revealed that DTCWT fusion is the best method under both JPEG and JPEG2000 compression, followed closely by CP fusion method. They also show that JPEG compression has a more severe effect on the fusion process, than JPEG2000. This is especially true at high compression ratios (over 1:50), where all the fusion methods seem to perform equally.
bad. These results are consistent with the psycho-visual experiments shown above.

Figure 10: IFQI for UN Camp image sequence

3.6 Discussion of Metric Results

A similar pattern of results was found between the more computationally advanced metrics and the general trends of the psychophysical results. The Petrovic and IFQI metrics both present the DT-CWT as performing the best, followed by the CP method, with the averaging method worst. In comparison with the results for $d'$ (Figures 3 and 5), the metric results fit well with the participants' performance. It should be noted that the psychophysical results for JPEG2000 did not find a significant compression effect, and this also seems to be reflected to a degree in the metric results, with a much less pronounced detriment in the results than JPEG, which did have a significant effect as well.

One reason for this could be that more highly compressed JPEG2000 images are required, as the point at which the JPEG2000 image detriment becomes significant could be more compressed than for JPEG. This is most apparent in the IFQI results, wherein a clear split is made between JPEG and JPEG2000 at low compression rates. The Petrovic metric, especially for AV, stays quite close between JPEG and JPEG2000. What this does mean, is that the metric results do not support the interaction found between the JPEG and JPEG2000 results with CP fusion, indicating that this fusion method cannot easily accommodate either method of fusion systematically. This is most probably due to the actual fusion process used by the CP method, which possibly heightened the contrast of the image as a whole, causing more parts of the image to become salient (e.g. the surrounding bushes etc.).

The second important result is that further evidence has been provided showing the JPEG2000 compression outperforms JPEG compression. Whilst other papers have focused more on quality assessment or simple mathematical comparisons, the current study provides both psychophysical and metric support for the use of the JPEG2000 standard, at least when performing certain tasks using fused images. Whilst the use of more advanced fusion methods does somewhat counteract the deficit caused to human performance, there is still a clear advantage in using the newer compression standard over the JPEG standard. This would allow users of image fusion to save vast amounts of computational storage space, especially as the current results show no worsening in human performance with higher levels of JPEG2000 compression. However, as posited in Section 3.4, a greater level of compression might be required for JPEG2000 to affect results, although previous research has already shown that JPEG2000 outperforms JPEG at higher compression levels [Steingrimsson and Simon 2003].

A final point of interest is the performance of the metrics; the widely used MI metric performed poorly when compared with human task performance, suggesting that the mutual information

4 Discussion

The current study set out to investigate how both fusion method and compression level might affect people’s perception of fused images. The novel use of a signal detection paradigm in this investigation, combined with computational metrics, has yielded some important findings, as well as paving the way for further investigation.

The first major finding is that both human task results and the more advanced computational metrics showed the DT-CWT scheme as providing the best performance, allowing participants to identify more accurately targets with almost no bias to answering in a particular direction. This is important in a field such as surveillance, where an individual might be asked to watch for a target over an extended period of time during which loss of concentration might cause the individual to become more biased in answering ‘no’. Likewise, if the image was to be fed into an automated monitoring, the use of the Petrovic or IFQI metrics to assess the quality of the fused image would seem to be appropriate according to the current findings.

These findings also suggest that the widely used CP method can underperform at times; this suggests that use of this method should be used with care. Moreover, the task-based results showed many variations for both JPEG and JPEG2000 results with CP fusion, indicating that this fusion method cannot easily accommodate either method of fusion systematically.
shared between the original input images and the fused image is not relevant in predicting human performance in the current scenario. However, the more advanced metrics, which use some aspects of the HVS, do seem to be providing closer results to those of the participants, in particular, the IFQI appears to be fairly robust in both JPEG and JPEG2000 results.

In conclusion, the current study has shown how a variety of image fusion and compression methods can be assessed using both objective human performance and metric performance. Further research is required in expanding the types of psychophysical tests used to find a more general picture. In addition, a wider range of fusion types could be used, and higher JPEG2000 compression levels could be adopted, to investigate where compression levels will affect task performance.

Acknowledgements

This work has been partially funded by the UK MOD Data and Information Fusion Defence Technology Center. The original ‘UN Camp’ infrared and visible images were supplied by Alexander Toet of the TNO Human Factors Research Institute. These images are available online at www.imagefusion.org.

References


