

Characterisation of Image Fusion Quality Metrics for Surveillance Applications over Bandlimited Channels

E. Fernández Canga, S. G. Nikolov,
C. N. Canagarajah and D. R. Bull
University of Bristol
Centre for Communications Research
Woodland Road, Bristol BS8 1UB
United Kingdom

E.Fernandez-Canga@bristol.ac.uk

T. D. Dixon, J. M. Noyes and T. Troscianko
University of Bristol
Department of Experimental Psychology
8 Woodland Road, Bristol BS8 1TN
United Kingdom
Timothy.Dixon@bristol.ac.uk

Abstract – Image fusion is finding increasing application in areas such as medical imaging, remote sensing or military surveillance using sensor networks. Many of these applications demand highly compressed data combined with error resilient coding due to the characteristics of the communication channel. In this respect, JPEG2000 has many advantages over previous image coding standards.

This paper evaluates and compares quality metrics for lossy compression using JPEG2000. Three representative image fusion algorithms: simple averaging, contrast pyramid and dual-tree complex wavelet transform based fusion have been considered. Numerous infrared and visible test images have been used. We compare these results with a psychophysical study where participants were asked to perform specific tasks and assess image fusion quality.

The results show that there is a correlation between most of the metrics and the psychophysical evaluation. They also indicate that selection of the correct fusion method has more impact on performance than the presence of compression.

Keywords: Image fusion, JPEG2000, quality metrics.

1 Introduction

Image fusion can be defined as the process of combining multiple input images into a smaller collection of images, usually a single one, which contains the relevant and important information from the inputs.

The aim of image fusion, apart from reducing the amount of data, is to create new images that are more suitable for the purposes of human/machine perception, and for further image-processing tasks such as segmentation, object detection or target recognition in applications such as remote sensing and medical imaging. Multi-sensor data often presents complementary information about region surveyed, scene or object, so image fusion provides an effective method to enable comparison and analysis of such data.

The benefits of multi-sensor image fusion include [1]: extended range of operation, extended spatial and temporal coverage, reduced uncertainty, increased reliability, robust system performance and compact representation of information.

The fusion process can take place at different levels of information representation, ranging from the earliest pixel level methods [2] to the more advanced region level algorithms [3]. Refer to [4] for a general review of fusion algorithms. Most of current image fusion applications use

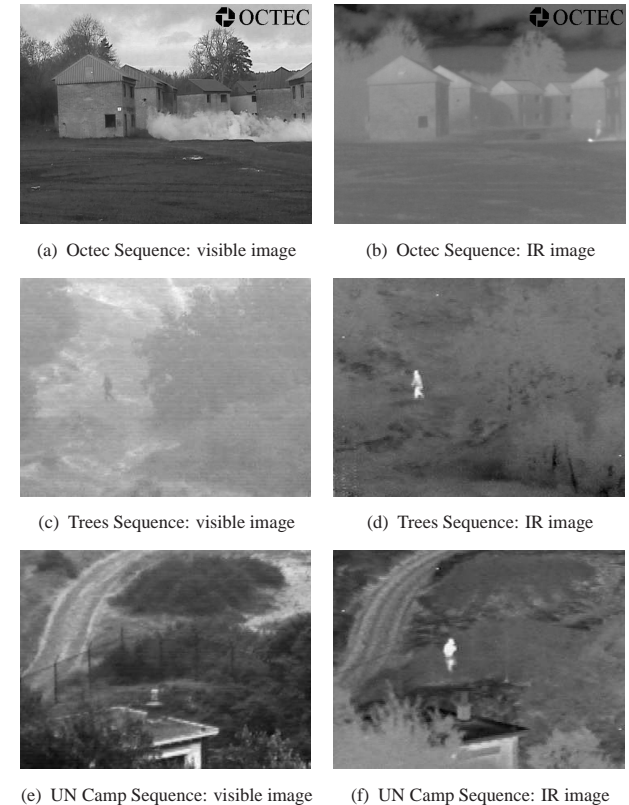


Fig. 1: Test image sequences

pixel-based methods, which are usually easy to implement and time efficient.

Image fusion applications such as remote sensing and military surveillance demand highly compressed data combined with error resilient coding due to the characteristics of the communication channel. In this respect, the JPEG2000 image compression standard [5] has many advantages over previous standards. JPEG2000 provides low bit-rate operation with rate distortion and subjective image quality performance superior to existing standards, without sacrificing performance at other points in the rate-distortion spectrum [6]. All these aspects have been rigorously tested and compared to previous compression standards [7, 8].

This paper studies the effects of compression on image fusion performance for surveillance applications. It evaluates and compares existing image quality metrics with

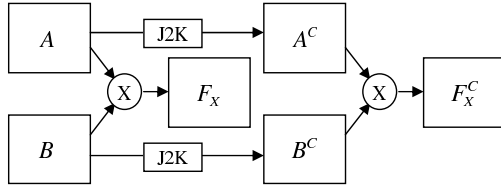


Fig. 2: Compression and fusion scheme

psycho-visual tests, where participants were required to perform specific tasks and to assess the visual quality of the fused images. In our experiments we use three representative image fusion algorithms: simple averaging (AV), contrast pyramid (CP) [9] fusion and dual-tree complex wavelet transform (DT-CWT) [10, 11] fusion. The images used are visible and infrared images from surveillance scenes, publicly available at [12], see Fig. 1.

The rest of the paper is organised as follows: Section 2 includes a short description of the test bed, Section 3 provides an overview of the metrics used in this paper, Section 4 details the psycho-visual experiments test bed. The results of both objective and subjective tests and their correlation are shown in Section 5. Finally, conclusions are drawn in Section 6.

2 Compression and Fusion

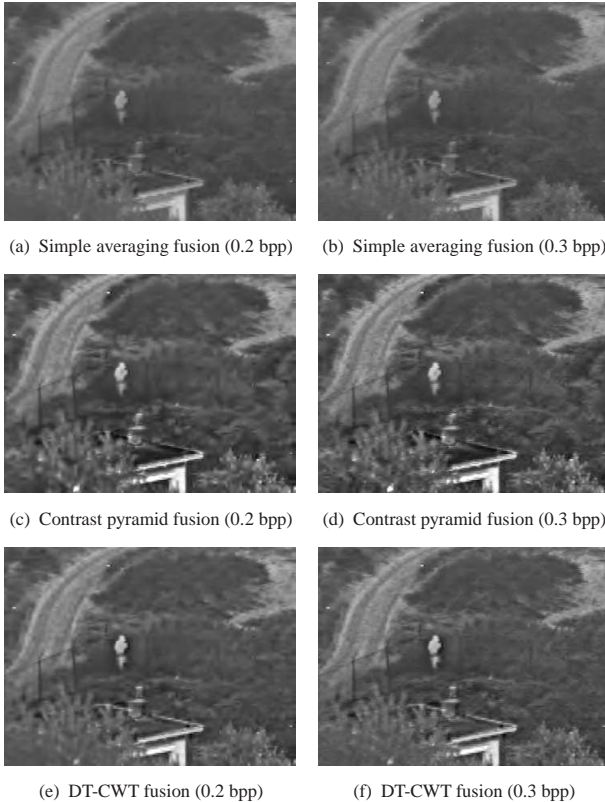


Fig. 3: Example of fused images

In order to measure the influence of compression on fusion performance, we generated multiple pairs of compressed images at different bit rates, from high compression ratios (1 : 100) to low compression ratios (1 : 5). Each

pair of images was then fused with the three fusion methods selected and the performance evaluated with the metrics described in Section 3.

The multi-resolution fusion methods (DT-CWT and CP) employed *four* levels of decomposition, maximum absolute value selection for high frequency coefficients and simple averaging for low frequency coefficients. Fig. 3 shows some examples of the fused images.

3 Metrics

Several approaches to fused image quality evaluation exist. These include qualitative tests with human participants and quantitative or objective tests. 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) [13], Petrovic and Xydeas metric [14, 15] and Piella's Quality Index [16] have been recently proposed. These estimate how and what information is transferred from the input images to the fused image.

3.1 Peak Signal to Noise Ratio

Peak signal to noise ratio is based on the root mean square error between the reference image R and the fused image F :

$$RMSE(R, F) = \sqrt{\frac{\sum_{i=1}^N \sum_{j=1}^M (R(i, j) - F(i, j))^2}{M \times N}} \quad (1)$$

where (i, j) denote pixel location. PSNR in decibels (dB) is then computed by using:

$$PSNR(R, F) = 20 \cdot \log\left(\frac{255}{RMSE(R, F)}\right) \quad (2)$$

Its main drawback is the assumption of knowledge of the ground-truth data, which is not feasible most of the time.

3.2 Mutual Information

Mutual information has emerged as an alternative to RMSE [13]. MI measures the degree of dependence of the two random variables A and B . It is defined by Kullback-Leibler measure:

$$I_{AB}(a, b) = \sum_{a, b} p_{AB}(a, b) \cdot \log \frac{p_{AB}(a, b)}{p_A(a) \cdot p_B(b)} \quad (3)$$

where $p_{AB}(a, b)$ is the joint distribution and $p_A(a) \cdot p_B(b)$ is the distribution associated with the case of complete independence.

Considering two input images A, B and a new fused image F , the amount of information that F contains about A and B can be calculated as:

$$I_{FA}(f, a) = \sum_{a, b} p_{FA}(f, a) \cdot \log \frac{p_{FA}(f, a)}{p_F(f) \cdot p_A(a)} \quad (4)$$

$$I_{FB}(f, b) = \sum_{a,b} p_{FB}(f, b) \cdot \log \frac{p_{FB}(f, b)}{p_F(f) \cdot p_B(b)} \quad (5)$$

and the image fusion performance measure can be defined as:

$$M_F^{AB} = \frac{I_{FA}(f, a) + I_{FB}(f, b)}{2} \quad (6)$$

3.3 Xydeas and Petrovic Metric

Recently, Petrovic and Xydeas [14, 15] 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.

It uses a Sobel edge operator to calculate the strength $g(n, m)$ and orientation $\alpha(n, m)$ information of each pixel in the input and output images.

The relative strength and orientation ‘change’ values, $G^{AF}(n, m)$ and $A^{AF}(n, m)$ of an input image A with respect to the fused image F , are defined as:

$$G^{AF}(n, m) = \begin{cases} \frac{g_F(n, m)}{g_A(n, m)}, & \text{if } g_A(n, m) > g_F(n, m) \\ \frac{g_A(n, m)}{g_F(n, m)}, & \text{otherwise} \end{cases} \quad (7)$$

$$A^{AF}(n, m) = \frac{|\alpha_A(n, m) - \alpha_F(n, m)| - \pi/2}{\pi/2} \quad (8)$$

These measures are then used to estimate the edge strength and orientation preservation values, $Q_g^{AF}(n, m)$ and $Q_\alpha^{AF}(n, m)$:

$$Q_g^{AF}(n, m) = \frac{\Gamma_g}{1 + e^{k_g(G^{AF}(n, m) - \sigma_g)}} \quad (9)$$

$$Q_\alpha^{AF}(n, m) = \frac{\Gamma_\alpha}{1 + e^{k_\alpha(A^{AF}(n, m) - \sigma_\alpha)}} \quad (10)$$

where the constants $\Gamma_g, k_g, \sigma_g, \Gamma_\alpha, k_\alpha, \sigma_\alpha$ determine the exact shape of the sigmoid nonlinearities used to form the edge strength and orientation. The overall edge information preservation values are then defined as:

$$Q^{AF}(n, m) = Q_g^{AF}(n, m) \cdot Q_\alpha^{AF}(n, m), \quad 0 \leq Q^{AF}(n, m) \leq 1 \quad (11)$$

A normalised weighted performance metric of a given process p that fuses A and B into F , is given as:

$$Q_p^{AB/F}(n, m) = \frac{\sum_{n=1}^N \sum_{m=1}^M Q^{AF}(n, m) w_A(n, m) + Q^{BF}(n, m) w_B(n, m)}{\sum_{n=1}^N \sum_{m=1}^M w_A(n, m) + w_B(n, m)} \quad (12)$$

It can be observed that the edge preservation values $Q^{AF}(n, m)$ and $Q^{BF}(n, m)$, are weighted by coefficients $w_A(n, m)$ and $w_B(n, m)$, which reflect the perceptual importance of the corresponding edge elements within the input images.

Note that in this method the visual information is associated with the *edge* information while the region information is ignored.

3.4 Image Fusion Quality Index

This image fusion quality index (IFQI) [16] is based on an image quality index recently introduced by Wang and Bovik [17], which is defined as:

$$Q = \frac{4\sigma_{xy}\bar{x}\bar{y}}{(\sigma_x^2 + \sigma_y^2)((\bar{x})^2 + (\bar{y})^2)} \quad (13)$$

where

$$\begin{aligned} \bar{x} &= \frac{1}{N} \sum_{i=1}^N x_i, \bar{y} = \frac{1}{N} \sum_{i=1}^N y_i \\ \sigma_x^2 &= \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2, \sigma_y^2 = \frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{y})^2 \\ \sigma_{xy} &= \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x}) \cdot (y_i - \bar{y}) \end{aligned}$$

To understand the meaning of Q , it can be decomposed as a product of three components:

$$Q = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \cdot \frac{2\bar{x}\bar{y}}{(\bar{x})^2 + (\bar{y})^2} \cdot \frac{2\sigma_x \sigma_y}{(\sigma_x^2 + \sigma_y^2)} \quad (14)$$

The first component is the correlation coefficient between x and y . The second component corresponds to the luminance distortion and the third factor measures the contrast distortion. The maximum value of $Q = 1$ is achieved when x and y are identical.

In order to apply this metric for image fusion evaluation, Piella and Heijmans [16] introduce salient information to the metric.

$$Q_w(A, B, F) = \sum_{w \in W} c(w) (\lambda(w) Q(A, F|w) + (1 - \lambda(w)) Q(B, F|w)) \quad (15)$$

where A and B are the input images, F is the fused image, $\lambda(w) = \frac{s(A|w)}{s(A|w) + s(B|w)}$ should reflect the relative importance of image A compared to image B within the window w , and $c(w)$ is the overall saliency of a window.

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 .

$$Q_E(A, B, F) = Q_w(A, B, F)^{1-\alpha} \cdot Q_w(A', B', F')^\alpha \quad (16)$$

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

As with the previous metrics, this metric does not require a ground-truth or reference image.

4 Task Performance Evaluation

The impetus behind using task-based image assessment methods comes from the lack of correlation found between subjective quality scores and computational metric performance [18]. It is therefore necessary to find a more suitable task with which to compare such metrics. The major benefit of fitting a task to the process of image fusion assessment is that this allows for accuracy ratings as well as response times to be recorded, thus, measuring the participants success. These objective measures then allow for a

more precise comparison with the results of computational metrics.

In choosing the most appropriate task for the current study, it was necessary to consider in what ways the original image sequences might have been used. It was initially decided that a range of tasks would be required in order to test the robustness of the metrics under varying conditions and constraints. This section covers the first experiment planned, which used the United Nations (UN) Camp Sequence (Figs. 1(e),1(f)), as used by Toet and colleagues [19] in two tasks to establish objective and subjective ratings of the images. These images comprise a sequence depicting a soldier moving around a UN Camp, which was shot using a visible light camera as well as an infrared camera.

The first part of the experiment used a backward masked rapid visual presentation paradigm, which involved watching a briefly presented image as in Fig. 3, with either the soldier present or absent. Participants had to state whether or not they thought the soldier had appeared in each frame viewed. In addition, subjective ratings were taken from participants in order to assess whether these would be comparable with the objective results. Subjective ratings were attained by presenting differing pairs of frames from the first part of the experiment, and asking participants to rate each on a scale of one to five, where one is thought to have 'very good' image quality, and five has 'very bad' quality.

4.1 Design

This experiment manipulated two independent variables in a related-measures design. The first variable was fusion type, with three types used: a simple averaging algorithm, a contrast pyramid method, and a dual-tree complex wavelet transform scheme. The second variable was JPEG 2000 compression level, also with three levels: clean (no compression), low (0.3 bpp) and high (0.2 bpp). The dependent variables were Hit rate (correctly identifying the soldier as present) and False Alarm rate (stating that the soldier was present when he was not). The trials were blocked by fusion type with compression type randomised within each block, and counterbalanced to avoid order effects. In each condition there were three different soldier-present images used with the soldier positioned approximately to the left of the image, in the centre of the image or to the right. Additionally, three soldier-absent images were used. Each image was displayed 10 times, thus creating a total of 540 trials, blocked into six blocks with 90 trials per block. A backward masking paradigm was used in order to stop ceiling effects by blocking potential afterimages when the test image was displayed, thus, disrupting further processing of the image [20].

In the second part of the experiment, image pairs were presented grouped by either fusion type (18 trials) or compression type (18 trials). Thus, one block showed pairs of images that were of the same fusion method, but were differing on compression level, whilst in the other block pairs were of the same compression but differed on fusion type. In both blocks, all combinations of fusion type and compression level were shown twice, so that each image was

shown on the left and right of the screen equal numbers of times. The two blocks were counterbalanced.

4.2 Participants

Twelve people participated in this experiment. These comprised of 10 females and two males, with a mean age of 20.25 years (range 18-26). Participants were required to have normal or corrected-to-normal vision to take part, and none of them had prior knowledge of the study.

4.3 Apparatus

Both parts of the experiment were displayed on a 19" flat screen CRT monitor. This was connected to a 2.8GHz Pentium 4 PC with 512 Megabytes RAM, running Superlab Pro v2.0 by Cedrus [21]. Responses were given using a regular keyboard. Written instructions were displayed at font size 30 in Arial script, and all presentations were centrally aligned. The test images were all monochrome, displayed at full size (360x270 pixels) against a 50% grey background, used in order to reduce harsh contrasting between the dark images and a white background.

The backward masks were created individually for each trial frame used by performing a Fourier transform on the image and randomising the phase component of the image, and then carrying out a reverse Fourier transform to create an image with equal power as the original but randomly distributed phase. A separate mask was created for each of the 540 trial images to ensure that any artefacts of one mask did not carry over to other masks.

In the second part of the experiment, the images were presented 70 pixels apart, with the full rating scale written out below the images in Arial size 24. The background surrounding the images in this case was white, in order to allow for a more absolute baseline assessment point. Each block used two images from the three frames used in the first part, both of which had the soldier present.

4.4 Procedure

After a consent form had been completed, participants were first shown the video sequence from which the four test frames were taken. This was shown once with the original visible light and infrared footage side by side, and again as a fused sequence with a brief explanation as to what image fusion was. It was made clear that the experiment would use frames from this sequence, and how an experimental trial would look.

In the first part of the experiment, participants were initially given 12 practice trials in which feedback was given as to whether they were correct or incorrect. Each trial in the practice and the main experiment began with a '+' fixation point appearing for 750ms in the centre of the screen, followed by the test image which was displayed for 15ms. There was then a 15ms Inter Stimulus Interval in which time the screen was blanked before the backward mask was presented for 250ms. The screen went blank, at which time the participants were required to press 'C' if the soldier was present and 'N' if he was absent. After every 90 trials there was a rest period.

In the second part of the experiment, it was explained that two images would be presented. Participants were required to give both images a rating from one to five, where one was deemed to have ‘very good’ image quality, and five has ‘very bad’ quality. On completing the experiment participants were thanked and debriefed.

4.5 Data Analysis

The scores for the first task were collated and Hit rates and False Alarm rates were obtained for each participant in each condition. These were used to calculate d' and beta values as in signal detection theory (e.g. [22]), with d' representing the distance between a signal and a noise response, and beta showing a participant’s criterion level for saying ‘present’ or ‘absent’. The d' score is calculated by calculating the Hit and False Alarm ratios for each condition. These are then converted into z-scores, which measure performance of a score in relation to the number of standard deviations it is below or above the mean. The d' sensitivity equals the Hit ratio z-score minus the False Alarm ratio z-score. The beta value is calculated by finding the ratio of Hit and False Alarm rate distribution curves at a given criterion. These values were then analysed using an Analysis of Variance (ANOVA) with the log of the beta being used, as a beta distribution is not statistically normal. For more information on the ANOVA statistical method see [23, 24].

5 Experimental Results

5.1 Psycho-Visual Results

The psycho-visual experiment was carried out in two parts, with the substantially longer objective test taking place first, followed by the subjective ratings task.

5.1.1 Part One of Experiment

The descriptive results for part one of the experiment showed that percent of hits varied across fusion type, with the averaging fusion method having a hit rate of 53%, whilst contrast pyramid had 79% and DT-CWT had 89% Hit rates. These did not seem to vary much across compression level with clean (73%), low (73%) and high (72%) compression scores. The False Alarm rates small differences between averaging (13%) and DT-CWT (13%) with contrast pyramid (10%), whilst clean (11%), low (13%) and high (11%) compression were also similarly matched.

Signal detection analysis of participants’ results showed that responses were closer to noise across the averaging fusion condition ($d' = 1.59$), less close with the contrast pyramid method (2.26), and furthest from noise with the DT-CWT (2.66). The results for the compression factor did not indicate any large differences, with clean (2.17), low compression (2.15) and high compression (2.20) being closely comparable.

Two factor repeated measure ANOVAs [24] carried out on these results backed up the general pattern detailed above. It was found that for d' , there was a significant main effect of fusion type ($F(2, 22) = 6.45, p = 0.006$), but not for compression level ($F(2, 22) = 0.18, p > 0.05$), nor was

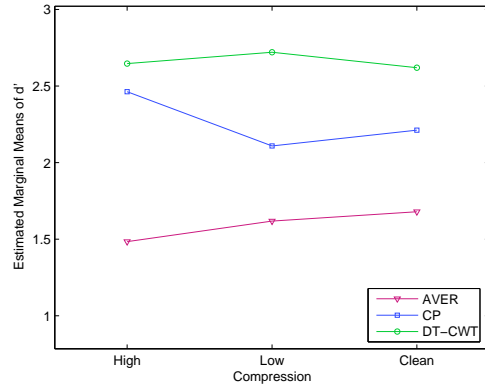


Fig. 4: ANOVA of d' comparing fusion and compression

there an interaction between the two factors ($F(4, 44) = 1.62, p > 0.05$), as shown in Fig. 4.

Post hoc testing of the pairwise comparisons using the Bonferroni test [23] showed that performance with the averaging method was not significantly lower than that with contrast pyramid (1.59 vs. 2.26, $p = 0.075$), although it is approaching significance on a two-tailed test. There was a significant difference between simple averaging and DT-CWT conditions (1.59 vs. 2.66, $p = 0.017$), but not between contrast pyramid and DT-CWT (2.26 vs. 2.66, $p > 0.05$). This indicates that the contrast pyramid and DT-CWT conditions had similar performance patterns to each other, both of which kept the signal and noise responses significantly further apart than the averaging condition.

Similarly, with the beta (bias) levels participants were much more likely to answer ‘no’ than yes in averaging condition ($beta = 1.37$), less so with contrast pyramid fusion (1.02), and unbiased either way with the DT-CWT (0.00). However, the compression levels again showed little difference ($clean = 0.82, low = 0.78, and high = 0.79$), indicating a regular, small bias towards giving a ‘no’ answer.

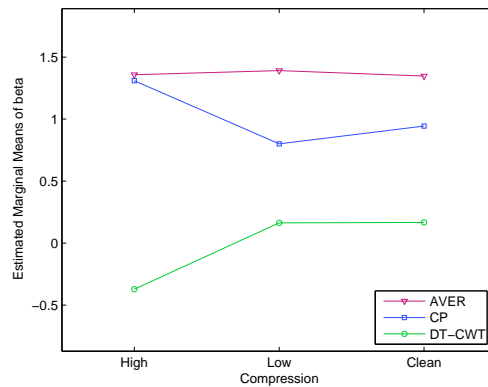


Fig. 5: ANOVA of beta comparing fusion and compression

ANOVAs showed that there was a significant main effect of fusion on beta ($F(2, 22) = 11.79, p < 0.001$), whilst there was no main effect of compression ($F(2, 22) = 0.073, p > 0.05$). However, in this case there was also an interaction found between the two factors ($F(4, 44) = 3.15, p = 0.023$), as shown in Fig. 5. Bonferroni tests revealed significant differences between averaging and DT-CWT (1.37 vs. 0.00, $p = 0.001$), and between contrast

pyramid and DT-CWT (1.02 vs 0.00, $p = 0.008$), but not between the averaging and contrast pyramid conditions (1.37 vs. 1.02, $p > 0.05$). This indicates that there is significantly less bias to answering ‘no’ in the DT-CWT condition than either of the other two conditions.

A Tukey’s Honestly Significant Difference test [24] was carried out to investigate the direction of the interaction. As seen in Fig. 5, there was a significant difference in the high compression condition between CP and DT-CWT fusion types (1.31 vs. -0.33 , $HSD = 0.84$, $p = 0.01$). This indicates that the high compression level was particularly affected by the difference between DT-CWT and CP fusion types, with participants much keener to answer ‘no’ when the high compression level and CP fusion was used than when high compression and DT-CWT fusion was used.

5.1.2 Discussion of Part One of the Experiment

It is clear that whilst fusion type was a critical factor in how the participants answered, compression type had little impact on the way they performed, with the one exception of the interaction covered above. What is of more interest is that there were different patterns of results between d' and beta within the fusion factor. Participants had significantly (or close to significantly) lower d' than both CP and DT-CWT fusion types, whilst these latter two fusion types did not differ significantly. However, beta scores showed that participants were significantly less biased with the DT-CWT fusion than for either contrast pyramid or averaging, with these latter two non-significant. Thus, it can be inferred that whilst the averaging fusion scheme creates a target that is significantly more confusable with noise than either CP or DT-CWT fusion, it is the DT-CWT method that leads to significantly lowered bias in participant answers, resulting in fewer ‘misses’ (saying ‘no’ when target is present) in answering.

5.1.3 Part Two of Experiment

The subjective scores of the participants were collated and the mean scores were considered. These showed that compression scores did vary somewhat, with clean (3.19) slightly higher than low compression (2.87) and high compression (2.68). Fusion scores appeared to vary more, although only for averaging fusion type, which scored much lower (2.34) than both contrast pyramid (3.16) and DT-CWT (3.25) methods.

A two-factor repeated measures ANOVA indicated a significant main effect of fusion type ($F(2, 22) = 22.87$, $p < 0.001$), as well as for compression level ($F(2, 22) = 11.80$, $p = 0.002$), but no interaction was found ($F(4, 44) = 0.65$, $p > 0.05$), as shown in Fig. 6. It should be noted that Mauchly’s test of sphericity was significant for both main effects, therefore the Greenhouse-Geisser test of within-subjects effects was used [23].

Post hoc testing using Bonferroni revealed significant differences between averaging and CP ratings (2.34 vs. 3.16, $p = 0.001$) as well, as between averaging and DT-CWT (2.34 vs. 3.25, $p = 0.001$), but not between CP and DT-CWT fusion methods (3.16 vs. 3.25, $p > 0.05$). This

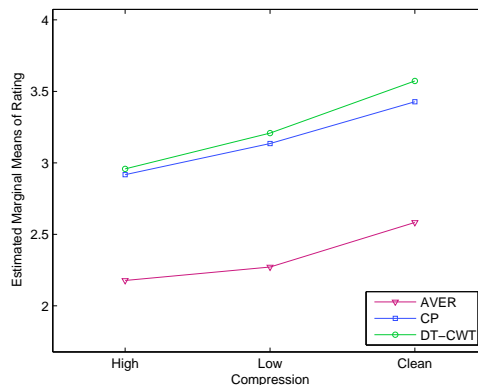


Fig. 6: ANOVA of ratings comparing fusion and compression

indicates that participants rated the subjective quality of the averaging fused images as lower than both the others, which were viewed as having a similarly higher quality.

Post hoc testing on the compression results revealed significant differences between clean and low compression images (3.19 vs. 2.87, $p = 0.045$), and between clean and high compression (3.19 vs. 2.68, $p = 0.008$), as well as between the low and high conditions (2.87 vs. 2.68, $p = 0.32$). These results indicate that participants rated the quality of the uncompressed image as the best of the three, followed by the less compressed image, with the highly compressed image scoring the lowest quality.

5.2 Discussion of Psycho-Visual Results

It is clear from the two psycho-visual sets of results that participants’ performance in objective tasks gives a similar pattern of results to how participants perceive subjective quality, although importantly, this is based on two factors not one. Performance in part one of the experiment hinged around fusion type, with the DT-CWT and CP fusion types leading to more separated signal-to-noise distributions, whilst averaging and CP fusion schemes were more likely to bias participants into making a ‘no’ response.

In contrast to the task-based results, participants’ subjective ratings were tiered both by compression and fusion type. The complex fusion method is preferred over the others, whilst a clean image is also rated more highly than the compressed images, as might be expected. What is interesting to note is that whilst compression level did not significantly affect the participants’ performances in the objective task at all, this was deemed to be as critical a factor as fusion type when the subjective quality was considered. Thus, compression is a factor that might be considered more important than it actually is if only subjective quality is considered. Dependent on what task is being carried out, this factor may only have minimal effect on performance. This general pattern of results shows the importance of measuring performance in carefully controlled experiments rather than relying on introspection or computational metrics alone, as these methods identify factors that might not be relevant when a task is being performed.

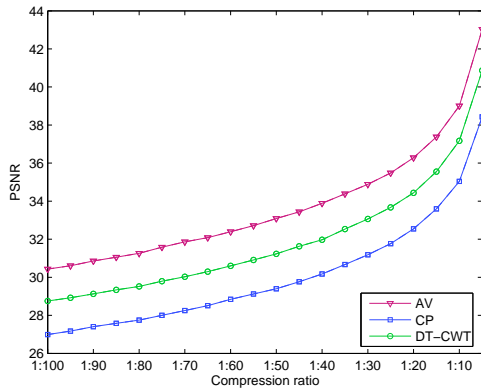


Fig. 7: PSNR for UN Camp image sequence

5.3 Metrics Results

The PSNR requires a reference or ideal image, which, in the case of IR and visible image fusion, is unknown most of the time. Therefore it is not really possible to assess the quality of fusion with this metric. However, it is possible to assess the influence of compression for each fusion method. In order to do so, we have obtained an uncompressed fused version (F_x) of the input images, where x represents the selected fusion method. We used this *clean* image as a reference to compare it with that obtained by fusing compressed images F_x^c (see Fig. 2).

Fig. 7 shows the PSNR measures for the three studied fusion methods at different compression rates. Examining this plot, the simple averaging fusion method seems to cope with compression better than the others. This is due to the low-pass filter characteristic of the averaging fusion method and the fact that the fusion rule in this case is linear. In the case of contrast pyramid and DT-CWT however, the fusion rule used was maximum absolute value, which is a non-linear operation. A small change in a pixel value can affect a number of coefficients in the transform domain. This may cause a switch in a number of coefficients selected by the fusion rule, which will affect a region in the image where the pixel values will change slightly.

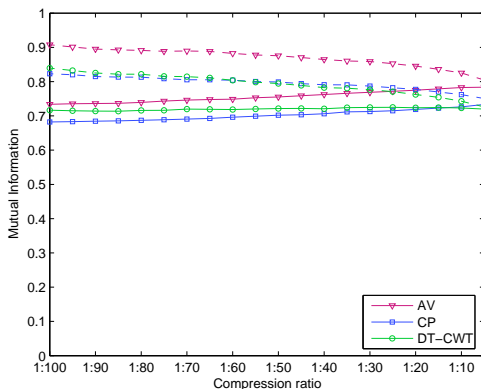


Fig. 8: Mutual information for UN Camp image sequence

Fig. 8 shows the results of mutual information measurements. The solid line represents the mutual information of uncompressed input images and the fused image obtained after compression ($M_{F_x^c}^{AB}$). The dashed line represents the

values obtained from the compressed input images and the corresponding fused image ($M_{F_x^c}^{A^c B^c}$).

The results obtained with this metric, contradict the ones obtained in the objective and subjective tasks. According to mutual information, averaging fusion method outperforms the other two methods and DT-CWT seems to be the least affected by compression.

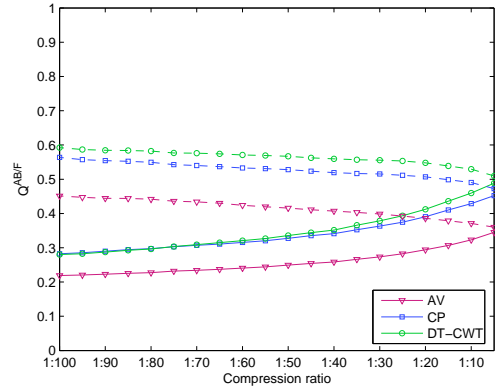


Fig. 9: Petrovic's metric for UN Camp image sequence

The results obtained by Petrovic and Xydeas' metric are presented in Fig. 9. It can be observed that the results correlate with the psycho-visual experiment, where the DT-CWT was the best fusion method. However, as compression increases (compression ratio of 1:70 or 0.11 bpp) the performance is very close to the contrast pyramid.

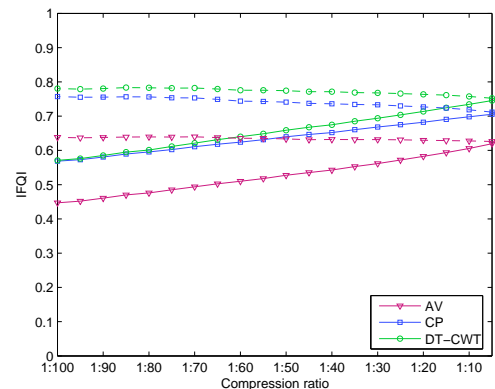


Fig. 10: Piella's metric for UN Camp image sequence

Similar results are obtained with Piella and Heijmans' metric using a window of size 8×8 and $\alpha = 0.2$ (Fig. 10). DT-CWT again performs best, with contrast pyramid in second position and averaging being the worst fusion method as expected. According to this metric, compression affects similarly the three fusion methods, with contrast pyramid having a slightly better performance.

6 Conclusions

In this paper we have discussed the influence of compression on image fusion. We reviewed commonly used quality metrics for image fusion and studied their performance with compressed images (JPEG2000). The results were compared with a psycho-visual study.

The performance of widely used quality metrics, such as mutual information, has been found to be poor. On the other hand, metrics that take into consideration aspects of the HVS, such as Petrovic and Xydeas' metric and Piella and Heijmans' metric seem to have a high correlation with subjective tests. Furthermore, these metrics together with the psycho-visual experiments show that the correct selection of the fusion method has a greater impact on image fusion performance than JPEG2000 compression itself.

Future work should include a study of the effects of transmission losses on image fusion performance.

Acknowledgements

This work has been partially funded by the UK MOD Data and Information Fusion Defence Technology Center. The original "UN Camp" and "Trees" IR and visible images are kindly supplied by Alexander Toet of the TNO Human Factors Research Institute and the Octec images by David Dwyer of OCTEC Ltd. These images are available online at www.imagefusion.org.

References

- [1] R. K. Sharma. *Probabilistic Model-based Multisensor Image Fusion*. PhD thesis, Oregon Graduate Institute of Science and Technology, Portland, Oregon, 1999.
- [2] P.J. Burt and R.J. Kolczynski. Enhanced image capture through fusion. *Fourth International Conference on Computer Vision*, pages 173 – 182, May 1993.
- [3] J. J. Lewis, R. J. O'Callaghan, S. G. Nikolov, D. R. Bull, and C. N. Canagarajah. Region-based image fusion using complex wavelets. In *Proceedings of the Seventh International Conference on Information Fusion*, volume I, pages 555–562. International Society of Information Fusion, June 2004.
- [4] G. Piella. A general framework for multiresolution image fusion: from pixels to regions. *Information Fusion*, 4(4):259 – 280, December 2003.
- [5] D. Taubman and M. Marcellin. *JPEG2000: Image Compression Fundamentals, Standards and Practice*. Kluwer International Series in Engineering and Computer Science. Kluwer Academic Publishers, 2002.
- [6] A. N. Skodras, C. A. Christopoulos, and T. Ebrahimi. Jpeg2000: The upcoming still image compression standard. *Pattern Recognition Letters*, 22(12):1337 – 1345, October 2001.
- [7] S. Battiato, A.R. Bruna, A. Buemi, and A. Castorina. Analysis and characterization of jpeg 2000 standard for imaging devices. *IEEE Transactions on Consumer Electronics*, 49(4):773 – 779, November 2003.
- [8] D. Santa-Cruz, R. Grosbois, and T. Ebrahimi. Jpeg 2000 performance evaluation and assessment. *Signal Processing: Image Communication*, 17(1):113–130, January 2002.
- [9] A. Toet. Multiscale contrast enhancement with applications to image fusion. *Optical Engineering*, 31:1026–1031, May 1992.
- [10] N. G. Kingsbury. The dual-tree complex wavelet transform: a new technique for shift invariance and directional filters. In *IEEE Digital Signal Processing Workshop*. 86, August, 1998.
- [11] S. Nikolov, P. Hill, D. Bull, and N. Canagarajah. Wavelets for image fusion. In A. Petrosian and F. Meyer, editors, *Wavelets in Signal and Image Analysis*, Computational Imaging and Vision Series, pages 213–244. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2001.
- [12] <http://www.imagefusion.org>.
- [13] Q. Guihong, Z. Dali, and Y. Pingfan. Medical image fusion by wavelet transform modulus maxima. *Optics Express*, 9(4):184, August 2001.
- [14] V. Petrovic and C. Xydeas. On the effects of sensor noise in pixel-level image fusion performance. *Proceedings of the Third International Conference on Information Fusion*, 2:14 – 19, July 2000.
- [15] V. Petrovic and C. Xydeas. Sensor noise effects on signal-level image fusion performance. *Information Fusion*, 4(3):167 – 183, September 2003.
- [16] G. Piella and H. A Heijmans. A new quality metric for image fusion. *International Conference on Image Processing, ICIP, Barcelona*, 2003.
- [17] Z. Wang and A.C Bovik. A universal image quality index. *Signal Processing Letters, IEEE*, 9(3):81 – 84, March 2002.
- [18] M. Miyahara, K. Kotani, and V.R. Algazi. Objective picture quality scale (pqs) for image coding. *IEEE Transactions on Communications*, 46(9):1215 – 1226, September 1998.
- [19] A. Toet, J. IJspeert, A. Waxman, and M. Aguilar. Fusion of visible and thermal imagery improves situational awareness. *Displays*, 18:85–95, 1997.
- [20] B. G. Breitmeyer and H. Ogmen. Recent models and findings in backward visual masking: A comparison, review, and update. *Perception & Psychophysics*, 62:1572 – 1595, 2000.
- [21] Superlab homepage. <http://www.superlab.com/>.
- [22] T. D. Wickens. *Elementary Signal Detection Theory*. Oxford University Press, 2002.
- [23] A. P. Field. *Discovering Statistics Using SPSS for Windows: Advanced Techniques for the Beginner*. SAGE Publications, London, 2000.
- [24] D. C. Howell. *Statistical Methods for Psychology*. Duxbury Press, Belmont, CA, 5 edition, 2001.