Perceptual Image Fusion Using Wavelets

Paul Hill, *Member, IEEE*, Mohammed Ebrahim Al-Mualla, *Senior Member, IEEE*, and David Bull, *Fellow, IEEE*

***Abstract*— A perceptual image fusion method is proposed that employs explicit luminance and contrast masking models. These models are combined to give the perceptual importance of each coefficient produced by the dual-tree complex wavelet transform of each input image. This combined model of perceptual importance is used to select which coefficients are retained and furthermore to determine how to present the retained information in the most effective way. This paper is the first to give a principled approach to image fusion from a perceptual perspective. Furthermore, the proposed method is shown to give improved quantitative and qualitative results compared with previously developed methods.**

***Index Terms*— Discrete wavelet transforms, image fusion, HVS.**

I. Introduction

THE effective fusion of two or more visual sources can provide significant benefits for visualisation, scene understanding, target recognition and situational awareness in multi-sensor applications such as medicine, surveillance and remote sensing.

The output of a fusion process should retain as much perceptually important information as possible from the two sources and should form a single more informative image (or video) [1], [2]. However, the majority of fusion methods described in the literature do not employ perceptual models of the Human Visual System (HVS) to decide which information to retain from each source. Furthermore, they do not suggest how to present this information so that it is perceived in the most effective way.

The lack of a perceptual basis for fusion is exemplified by the use of transform techniques that exploit multiscale decompositions such as wavelets. These approaches assume an implicit linear relationship between the magnitude of a transform coefficient and its significance, without any reference to perceptual models within the given transform domain. In contrast, our work utilises state of the art models of perceptual significance, previously developed for the effective compression of video and image content. These models characterise perceptual significance taking into account of: i) luminance masking, ii) the variation of contrast perception with frequency using Contrast Sensitivity Functions (CSFs) and iii) contrast masking.

Previously developed models have been based on critically decimated transforms such as the Discrete Cosine Transform (DCT) or the Discrete Wavelet Transform (DWT). We have generalised these models for use with a multiscale decomposition more effective for fusion: The Dual Tree Complex Wavelet Transform (DT-CWT) [1], [3].

The remainder of this paper is organised as follows. Firstly, in section II the context of this work is described together with the challenges faced together with relevant previous work. Secondly, perceptual models for image fusion are developed within section II-D. These models together with relevant fusion rules are further described in section IV. Fusion results are then described in section V and finally conclusions are presented within section VI.

Manuscript received February 25, 2016; revised July 18, 2016 and October 13, 2016; accepted November 12, 2016. Date of publication November 30, 2016; date of current version January 20, 2017. This work was supported by EPSRC Platform under Grant EP/M000885/1. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Catarina Brites.

P. Hill and D. Bull are with the Department of Electrical and Electronic Engineering, University of Bristol, Bristol BS8 1UB, U.K. (e-mail: paul.hill@bristol.ac.uk).

M. E. Al-Mualla is with the Khalifa University of Science, Technology & Research, Abu Dhabi 127788, United Arab Emirates.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2016.2633863

II. Challenges and Review

Image fusion can exploit a variety of pixel or transform domain methods to combine important or salient information from two or more images into a single fused output. Multiscale transforms have been shown to provide good fusion performance [4]. Wavelet transforms in particular provide a flexible multiscale fusion structure within a well understood mathematical framework [1], [2], [5]. We have therefore adopted wavelet transforms for our work and have combined them with a perceptual significance model.

A. Discrete Wavelet Transform (DWT) Based Image Fusion

Early fusion methods based on pyramid decompositions [6] have now largely been superseded by DWT-based methods [1], [2]. The fusion of two sources utilising the Discrete Wavelet Transform can be defined in terms of the two registered input sources $\boldsymbol{S^0}$ and $\boldsymbol{S^1}$, the wavelet transform itself $\boldsymbol{\omega}$ and a fusion rule $\boldsymbol{\phi}$, defined to combine co-located coefficients within the transform domain. The fused wavelet coefficients are then inverted using an inverse wavelet transform $\boldsymbol{\omega^{-1}}$ to produce the resulting fused image $\boldsymbol{F}$, thus:

$$\boldsymbol{F} = \boldsymbol{\omega^{-1}}(\boldsymbol{\phi}(\boldsymbol{\omega}(\boldsymbol{S^0}), \boldsymbol{\omega}(\boldsymbol{S^1}))). \tag{1}$$

B. Dual Tree Complex Wavelet Transform (DT-CWT) Based Image Fusion

Although DWT-based image fusion provides good results, it has been recognised that the associated shift variance produces sub-optimal performance. Specifically this is because a coefficient's magnitude may not accurately reflect the actual energy attributable to its spatial-frequency location. It is therefore an

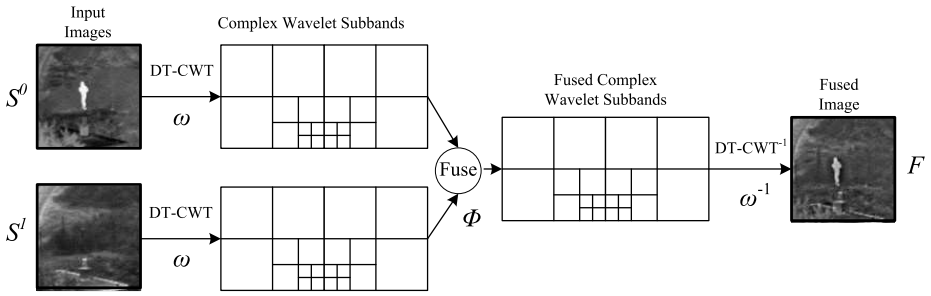

Fig. 1. Fusion of two images using the DT-CWT.

inaccurate basis for comparing perceptual importance within the transform domain [1], [3].

This problem has been addressed through the use of transforms such as the Shift Invariant Discrete Wavelet Transform (SI-DWT) [5]. This transform is identical to the DWT but removes the down-sampling at each stage of the decomposition. Since the shift variation is caused by down-sampling, the SI-DWT is shift invariant.

The Dual Tree Complex Wavelet Transform (DT-CWT) provides a significantly more compact transform domain representation that not only achieves near shift-invariance but also provides improved directionality (through the ability to distinguish positive and negative frequencies) [3]. This discrimination (of positive and negative frequencies) results in the division of each of the original three DWT subbands to produce six DT-CWT subbands, centred at orientations $\pm 15°$ $\pm 45°$ and $\pm 75°$. The resulting subband coefficients are complex and the magnitude of each coefficient is approximately shift invariant. Coefficient magnitudes can therefore be effectively used to fuse two transformed images. A typical fusion scenario is illustrated in figure 1.

The DT-CWT offers several advantages for image fusion over other transform based methods. Firstly it is considerably less over-complete than the SI-DWT. Secondly, the improved directional selectivity and reduced shift variance results in improved fusion performance compared to the DWT [1], [7].

C. Perceptually Based Image Fusion

A small number of researchers have attempted to integrate perceptual criteria into image fusion applications. Nercessian *et al.* [8] proposed a method based on a Laplacian pyramid decomposition using fusion rules dependent on a simple local Weber law relationship. A perceptually based image fusion method using the Laplacian pyramid decomposition was also developed by Mertens *et al.* [46]. Li *et al.* [9], Bhatnagar and Liu [10], and Li *et al.* [11] have all integrated a localised version of a "visibility metric" (developed by Huang *et al.* [12] into the fusion process. Although these papers claim to offer HVS-based fusion, the visibility metric used is simply a weighted local variance and therefore is not specifically dependent on luminance and contrast masking. Wang and Ye [13] presented a Weber-law based luminance adaptation method integrated within a "total variation" fusion approach. This uses a global luminance adaptation function (i.e. the global threshold line shown in figure 2). We are not aware of any prior art on integrating contrast masking into an image fusion application.

D. Perceptual Models for Image Fusion

A fundamental finding of psychovisual and physiological investigations into the human visual system is that the eye's photoreceptors and the connected neurons interact in complex ways throughout the early stages of vision. For example, receptive fields surrounding each ganglion cell within the retina have been found to contain both excitatory and inhibitory responses [14]. The perception of contrast relevant to image fusion has therefore been found to be dependent on a range of masking effects. Three relevant masking effects employed in our work are:

1) Luminance Masking: The dependence of contrast perception on local luminance.
2) Contrast Masking: Contrast perception is also dependent on orientation of local content.
3) Frequency Masking: The Contrast Sensitivity Function (CSF) gives a measure of the perceptual importance of spatial frequencies.

Numerous models have been proposed to predict the behaviour of the HVS for each of these masking effects, both individually and collectively. For example, comprehensive models of luminance contrast have been defined by Daly [15] and Barten [16]. Additionally, many complete visual models incorporating these types of masking to predict the perceptual significance of any pattern have been implemented. These include:

- The visual predictor defined by Daly [15] utilising the Cortex transform.
- The foveal detection model of local contrast defined by Watson and Ahumada [17].
- The High Dynamic Range (HDR) predictor, HDR-VDP-2 defined by Mantiuk *et al.* [18] using the steerable pyramid transform.

Unfortunately, however, these "difference prediction" perceptual models are not directly applicable to image fusion. These predictors process two input images to produce a localised perceptual difference. Although these output difference measures exploit similar perceptual models to those used in this paper, the output represents the difference between the images rather than their relative perceptual importance.

To overcome this, we have adopted the perceptual model framework proposed for compression applications by Höntsch and Karam [19] and adopted by Liu *et al.* [20].

III. A Framework for Perceptual Fusion

Our model utilises the ratio of a coefficient's magnitude to the locally calculated Just Noticeable Difference (JND) threshold. We base the calculation of JND on the model proposed by Liu *et al.* [20] where the local JND associated with a coefficient (at spatial position i,j) is defined as:

$$t_{JND}(\lambda,\theta,i,j) = JND_{\lambda,\theta}a_l(\lambda,\theta,i,j)a_c(\lambda,\theta,i,j), \quad (2)$$

where $JND_{\lambda,\theta}$ is the baseline frequency-dependent contrast sensitivity JND threshold according the central frequency of subband scale λ at orientation θ; $a_l(\lambda,\theta,i,j)$ is the luminance masking effect and $a_c(\lambda,\theta,i,j)$ is the contrast masking effect.

For all the fusion rules, only co-located coefficients within the same subbands are compared. Therefore $JND_{\lambda,\theta}$ will be constant for both images' coefficients. We define our local JND threshold as:

$$t_{JND}(\lambda,\theta,i,j) = a_l(\lambda,\theta,i,j)a_c(\lambda,\theta,i,j). \quad (3)$$

Although the contrast sensitivity JND threshold $JND_{\lambda,\theta}$ is not used in (3), it is utilised within the intra and inter-band methods described in section III-B.

A. Luminance Masking/Adaptation

The perception of contrast by the human visual system is dependent on the luminance context i.e. on the global and local background luminance levels. This is termed luminance adaptation and has conventionally been represented using the Weber-Fechner law or its power law variants. This has been modelled using the Ahumada-Peterson and DCTune formulae for perceptually based DCT and wavelet coefficient quantisation and JND threshold calculation in compression applications [20]–[22].

The Weber-Fechner law states that the ratio of the JND threshold T to the background luminance L is constant over a range of L.[1] Modified versions of this law have been utilised by Ma and Huang [23] and Wang and Ye [13] where the ratio T/L is elevated for high and low background luminance values to take into account the human visual system's decrease in sensitivity within these background luminance regions.[2]

More recent models for luminance adaptation have been based on the observation that the Weber-Fechner law is an over simplification for typical viewing conditions. Specifically, it has been suggested that luminance based JND thresholds vary as a "quasi-parabola" or u-shaped curve with reference to a more global background luminance average. It has been proposed that the shape of this "quasi-parabolic" curve is attributable to either the local nature of the luminance background [24] or to the fact that the gamma correction of typical monitors will lead to higher thresholds at lower luminance values [25], [26]. For either reason, such u-shaped models have formed the basis of the majority of recent luminance adaptation models [24]–[32] and has been successfully applied in compression applications, especially for high dynamic range video coding [33]. Figure 2 illustrates how these u-shaped local background curves are modulated by an overall Weber-Fechner type global luminance masking effect.

[1]This is illustrated by the middle section of the main graph in figure 2

[2]This is illustrated by the overall shape of the main graph in figure 2

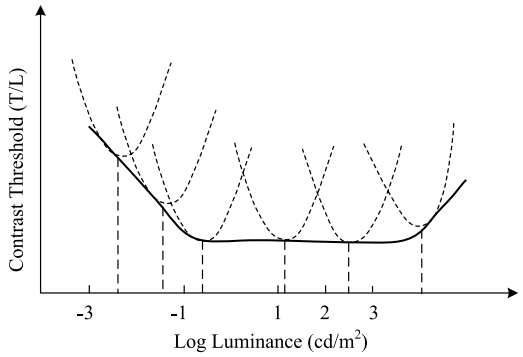

Fig. 2. Local and global luminance JND threshold models [25].

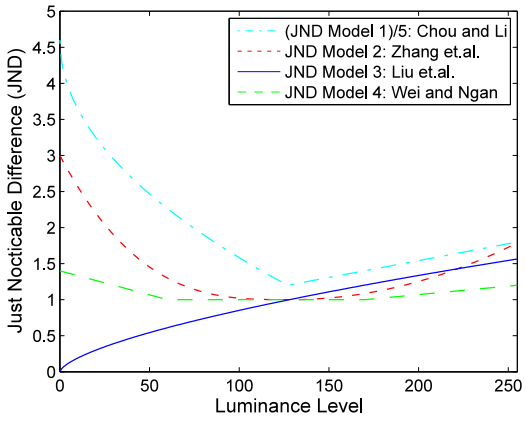

Fig. 3. Luminance Masking/Adaptation (Just Noticeable Difference) Models.

1) Luminance Masking Model Definitions: Figure 3 shows four example luminance adaptation models that are defined in (4), (5), (6) and (7). These models define the just noticeable difference (a_l) attributable to variations in the local background luminance average $\overline{I}$.

Luminance Adaptation Model 1: Proposed by Chou and Li [28] and used by Lee *et al.* [29] and Yang *et al.* [30].

$$a_l = \begin{cases} 17\left(1-\sqrt{\dfrac{\overline{I}}{127}}\right)+3, & \text{if } \overline{I} \le 127 \\ \dfrac{3}{128}(\overline{I}-127)+3, & \text{otherwise.} \end{cases} \quad (4)$$

This was the model selected to calculate (3) defined above.

Luminance Adaptation Model 2: Proposed by Zhang *et al.* [25] utilised by Zhang *et al.* [27].

$$a_l = \begin{cases} 2\left(1-\dfrac{\overline{I}}{128}\right)^3+1, & \text{if } \overline{I} \le 128 \\ 0.8\left(\dfrac{\overline{I}}{128}-1\right)^2+1, & \text{otherwise.} \end{cases} \quad (5)$$

Luminance Adaptation Model 3: Proposed by Ahumada and Peterson [21] used by Watson *et al.* [22]

and Liu *et al.* [20].

$$a_l = \left(\frac{\overline{I}}{128} \right)^{0.649}. \quad (6)$$

Luminance Adaptation Model 4: Wei and Ngan [26].

$$a_l = \begin{cases} \left(60 - \overline{I}\right)/150 + 1, & \text{if } \overline{I} \leq 60 \\ 1, & \text{if } 60 < \overline{I} < 170 \\ \left(\overline{I} - 170\right)/425 + 1, & \text{if } \overline{I} \geq 170 \end{cases} \quad (7)$$

Figure 3 shows that models 1, 2 and 4 are u-shaped whereas model 3 is monotonically increasing over the background luminance range. It should be noted that this figure plots the first model scaled by one fifth so that it can be easily compared to the other models. This large discrepancy is not explained by the authors of the method [28]. However, since most applications require relative JND comparisons, its effect is not thought to be significant. The threshold values of $\overline{I}$ above 128 in model 2 were designed to approximate model 3 for this range. This approximation is also illustrated in figure 3.

Alternative models include the modified Weber-Fechner model proposed by Ma and Huang [23] and used by Wang and Ye [13]. This model was not investigated in our work because its simplistic form does not reflect the types of application and viewing conditions relevant to image fusion. An additional "quasi parabola" type model was proposed by Safranek and Johnston [31] and referenced by Miloslavski and Ho [32]. This model was also not used as it was not precisely defined in these papers.

Luminance adaptation model 1 was selected to calculate (3) as it gave the best results (see section V).

2) Calculation of Local Luminance $\overline{I}$: In DCT based compression models, $\overline{I}$ is defined as the DC value of the transform block (or average of a group of local DC values). For DWT based compression models, $\overline{I}$ is calculated as the co-located coefficient of the Low-Low subband (relative to the considered high-pass coefficient). Ideally, this should be calculated with reference to the foveal region and is therefore dependent on viewing distance and screen resolution. However, in most cases, the variation (with viewing distance and/or screen resolution) is not considered significant. Therefore, a fixed number of local pixels (such as that produced by a set number of DWT decomposition levels or the DCT block area) is assumed.

In our model, we have calculated $\overline{I}$ from the magnitude of the co-located lowpass coefficient at the highest decomposition level (adjusted to have the same dynamic range as the image).

It should be noted that all the above models assume a fixed relationship between all values of $\overline{I}$ and the actual luminance. However, this is the assumption of all the previously cited work using the generalisation model illustrated in figure 2.

B. Contrast Masking

Contrast masking quantifies how the visibility of an image component (the target) varies in the presence of other image components (the masker) [34]–[36]. Contrast masking is commonly defined as the variation of the JND threshold of a target signal as a function of the intensity of a masking (or masker) signal [34]. Within a fusion framework, the target signal can be defined as a transform coefficient and the masking signals as the neighbouring coefficients (i.e. spatially neighbouring coefficients and local coefficients in neighbouring frequency subbands).

The measure of contrast masking a_c can be defined and modelled using the DT-CWT as:

$$a_c(\lambda, \theta, i, j) = a_{c\_intra}(\lambda, \theta, i, j) a_{c\_inter}(\lambda, \theta, i, j), \quad (8)$$

where $a_{c\_intra}(\lambda, \theta, i, j)$ is the contrast masking effect due to the coefficients within the same subband as the "target" coefficient and $a_{c\_inter}(\lambda, \theta, i, j)$ is the contrast masking effect due to the co-located coefficients within neighbouring subbands ((λ, θ, i, j) is the subband and spatial location within this subband as defined above).

Our contrast masking model is based on that used for wavelet coding by Höntsch and Karam [19] and implemented by Liu *et al.* [20]. It also relates very closely to the model proposed by Mantiuk *et al.* [18] for high dynamic range difference predictions. This model also takes into account intraband and interband masking within a steerable pyramid decomposition. Our model differs from these in that it explicitly weights the variation of masking with orientation (according to the results of Foley [34]) and takes into account a larger number of more closely related coefficients (in orientation, frequency and location). It also uses a more recently defined weighting of neighbouring subbands' CSF [37].

1) Intraband Contrast Masking: Watson [38] reported that contrast masking has the largest effect when the target and masker have the same frequency and orientation. Hence intraband contrast masking is considered separately from interband masking.

We now define intraband contrast masking for coefficient $v_{\lambda,\theta,i,j}$ (i.e. position (i, j) within subband (λ, θ)). As (λ, θ) is constant within the same subband we simplify the notation of $v_{\lambda,\theta,i,j}$ to be $v_{i,j}$. This intraband contrast masking is based on a weighted version of the contrast masking defined by Höntsch and Karam [19] and implemented by Liu *et al.* [20]. It is defined as:

$$a_{c\_intra}(\lambda, \theta, i, j) = \max \left\{ 1, W_{intra} \sum_{v \in C_{i,j}(h)} \left| \frac{v}{JND_{\lambda,\theta}} \right|^{\zeta} / N_{i,j} \right\}, \quad (9)$$

where $C_{i,j}(h)$ is the "deleted neighbourhood set" of size $(2h + 1) \times (2h + 1)$ centred around coefficient $v_{i,j}$ within subband (λ, θ). i.e. the set within a square window minus the centre:

$$C_{i,j}(h) = \{v_{i+m,j+n} | -h \leq m, n \leq h\} \setminus \{v_{i,j}\}. \quad (10)$$

The weighting factor W_{intra} and exponent factor ζ are defined to weight the effect of intra masking (set to 12 and 0.6 respectively). $JND_{\lambda,\theta}$ is the contrast sensitivity threshold of subband (λ, θ) defined in section III-B.3. $N_{i,j}$ is the number of coefficients in $C_{ij}(h)$.

2) Interband Contrast Masking: Many contrast masking models only use intraband masking. While this offers a conservative and tractable model (for example Liu *et al.* [20]), previous vision research gives more support to models with a wider spatial, orientation and frequency spread of the masking signal [34], [39].

We therefore define such an interband model and define interband contrast masking for coefficient $v_{\lambda,\theta,i,j}$. As the position (i, j) is constant for all the co-located coefficients, we simplify the notation of $v_{\lambda,\theta,i,j}$ to be $v_{\lambda,\theta}$. As with the intraband masking, interband contrast masking is based on a weighted version of contrast masking defined by Höntsch and Karam [19] (and related to the intraband contrast masking implemented by Liu *et al.* [20]). Intraband contrast masking for a coefficient for subband (λ, θ) at position (i, j) is defined as:

$$a_{c\_inter}(\lambda, \theta, i, j) = \max\left\{1, W_{inter} \sum_{v \in C_{\lambda,\theta}} w_\lambda w_\theta \left|\frac{v}{JND_{\lambda,\theta}}\right|^\beta / N_{i,j}\right\}, \quad (11)$$

where $C_{\lambda,\theta}$ contains the co-located coefficients $v_{\lambda,\theta}$ from all the orientated subbands at the same scale as the target coefficient together with all the orientated subbands at the neighbouring scales $\lambda - 1$ and $\lambda + 1$ (minus the subband coefficient at the same scale and orientation at the target) i.e.

$$C_{\lambda,\theta} = \{v_{\lambda+d,\theta_e} | d = -1, 0, 1, \theta_e = \pm 15°, \pm 45°, \pm 75°\} \setminus \{v_{\lambda,\theta}\}. \quad (12)$$

The weighting factor W_{inter} and exponent factor β are defined to weight the effect of interband masking (set to 12 and 0.6 respectively). $JND_{\lambda,\theta}$ is the JND contrast sensitivity threshold of subband (λ, θ) defined in section III-B.3. $N_{i,j}$ is the number of coefficients in $C_{\lambda,\theta}$.

The effect of contrast masking is only observed up to a relative frequency range of two octaves [14]. We have therefore only taken into account neighbouring scales within the DT-CWT as this represents a conservative model. Additionally, masking experiments conducted by Foley [34] have indicated that there is an approximately linear decrease in the masking effect as the orientation difference between a target and masking signal increases. This results in a maximum decrease of approximately 10:1. The orientation weights w_θ indicating the masking effect of an orientated masking subband to the target subband have therefore been defined as:

$$w_\theta = \begin{cases} 1.0 & \angle\theta_T\theta_M = 0° \\ 0.7 & \angle\theta_T\theta_M = \pm 30° \\ 0.4 & \angle\theta_T\theta_M = \pm 60° \\ 0.1 & \angle\theta_T\theta_M = \pm 90°, \end{cases} \quad (13)$$

where $\angle\theta_T\theta_M$ is the angle between the target subband and the masking subband.

$$w_\lambda = \begin{cases} 1.0 & \lambda_T - \lambda_M = 0 \\ 0.5 & \lambda_T - \lambda_M = \pm 1 \\ 0 & \text{otherwise.} \end{cases} \quad (14)$$

TABLE I
BASIS FUNCTION AMPLITUDE $A_{\theta,\lambda}$ FOR A SIX-LEVEL DT-CWT

	DT-CWT Decomposition Level, λ					
Orien θ	1	2	3	4	5	6
$\pm 15°$, $\pm 75°$	0.1684	0.1481	0.0704	0.0347	0.0169	0.0082
$\pm 45°$	0.1684	0.1583	0.0767	0.0395	0.0202	0.0100

TABLE II
PARAMETERS FOR THRESHOLD MODEL (16)

Transform	a	k	f_0	g_1	g_2	g_3
DWT (from [41])	0.495	0.466	0.401	1	0.534	1
DT-CWT [37]	3.107	1.025	0.755	1	0.814	1

The interscale weight w_λ is defined as being 1 for the same subband and 0.5 for neighbouring scales. For example if the target subband is aligned to 15°, the masking subband is aligned to 75° and the subbands are separated by one scale then the total intraband contrast weight $w_\lambda w_\theta = 0.4 \times 0.5 = 0.2$.

3) Contrast Sensitivity: The contrast sensitivity measure $JND_{\lambda,\theta}$ used in (9) and (11) defines the JND threshold for a given spatial frequency. This JND threshold is the inverse of the contrast sensitivity function which has a maximum at around 2-5 cycles per degree [40] and has an approximately parabolic shape. These functions are conventionally created by finding the threshold of perception of Gabor functions when varying their spatial frequency [41]. However, for the case of wavelet transforms, the spatial support of the basis functions varies with frequency. Watson *et al.* [41] conducted JND experiments with discrete wavelet transform basis functions and found that the JND threshold function monotonically increases for the considered spatial frequencies. This function was modelled in [41] to be parabolic in the log-log (threshold-frequency) domain as given by the parametric equation in (15):

$$log_{10}(Y) = log_{10}a + k(log_{10}f - log_{10}(g_\theta f_0))^2, \quad (15)$$

where Y is the luminance JND, f is the frequency and the remaining parameters are defined in [41] and below. This is related to the JND of each basis function (16) as:

$$JND_{\lambda,\theta}(r) = \frac{1}{A_{\lambda,\theta}} a 10^{k\left\{\log_{10}\left(\frac{g_\theta f_0 2^\lambda}{r}\right)\right\}^2}, \quad (16)$$

where $A_{\lambda,\theta}$ is the amplitude of the basis function and r is the visual resolution of the used display in pixels per degree, which can be calculated as:

$$r = dw \tan\left(\frac{\pi}{180}\right) \approx d\frac{w\pi}{180} \approx d\frac{w}{57.3}, \quad (17)$$

where w is the viewing distance (in cm) and d is the display resolution in pixels/cm.

Additionally, Hill *et al.* [37] used similar experiments for the DT-CWT and obtained the parameters of (16) for the basis functions of the DT-CWT as shown in table II. Furthermore, the amplitude $A_{\lambda,\theta}$ for DT-CWT basis functions used is shown in table I. The viewing distance w used for

these experiments was approximately 57cm. Therefore for the experiments within [37] $r \approx d$ leading to a display visual resolution r of 21.38 pixels/degree (10.69 cycles/degree).

Although different viewing conditions will alter the threshold function, the monotonically increasing nature of the function will enable the differences between two scales (as used in (9) and (11)) to generalise more effectively.

IV. Perceptually Based Fusion Rules

A. DT-CWT Coefficient Fusion Decision

Figure 1 shows the structure of how DT-CWT coefficients from two images can be combined to form a fused transform that, when inverse transformed, forms the fused image. The simplest form of a wavelet coefficient fusion rule is the *choose maximum* rule (as implemented for the DWT in [2] and the DT-CWT in [1]). The choose maximum fusion rule, or its variants, makes the assumption that the perceptual importance of a coefficient is directly related to its magnitude.

Our method for choosing the most perceptually important coefficient is as follows. Firstly we obtain the measure of perceptual importance or "Noticeability Index" of each coefficient within each image. For each coefficient we determine its perceptual importance as the ratio of its magnitude to the locally calculated JND threshold t_{JND}[3]:

$$NI^I(\lambda,\theta,i,j) = \frac{\left|v^I_{\lambda,\theta,i,j}\right|}{t^I_{JND}(\lambda,\theta,i,j)}, \tag{18}$$

where NI is the "Noticeability Index", $I \in \{0,1\}$ is the index of the images to be fused, $v^I_{\lambda,\theta,i,j}$ is the DT-CWT coefficient within subband λ,θ at spatial position i,j and $t^I_{JND}(\lambda,\theta,i,j)$ is the local JND calculated by (3). Luminance adaptation model 1 is selected to calculate (3). The choice of coefficient from the two images is based on the largest NI (dropping the indexes λ,θ,i,j for clarity):

$$v^{Max(NI)} = \begin{cases} v^0 & NI^0 > NI^1 \\ v^1 & \text{otherwise.} \end{cases} \tag{19}$$

Alternative fusion rules that combine coefficients by considering local salient regions [2] or using weighted mixtures of coefficients between the two images [43] have also been reported elsewhere. However, these schemes do not have any explicit perceptual basis.

B. Perceptually Fused Coefficients

To render perceptually important content from each image, the most perceptually important coefficient (chosen using (19)) will need to be adjusted. This ensures that the output coefficient retains the same perceptual importance in the fused image as it did in its original image (according to our models). Continuing to assume a linear "noticeability" of coefficient magnitudes relative to t_{JND}, each coefficient can be adjusted as follows:

$$v^3 = \begin{cases} v^0 \dfrac{t^2_{JND}}{t^0_{JND}} & NI^0 > NI^1, \\ v^1 \dfrac{t^2_{JND}}{t^1_{JND}} & \text{otherwise,} \end{cases} \tag{20}$$

where v^3 is the fused coefficient (identified by the spatial and subband indices $\{\lambda,\theta,i,j\}$) for each subband at each spatial location and t^2_{JND} is the local JND threshold for the fused image. As the fused image has not been created at this point t^2_{JND} is calculated from a pre-fused, non-perceptually based *choose maximum* fused image. This pre-fused image will have representative luminance and contrast content to accurately offset the final fused coefficient so it will be perceived with the same visual importance as when it was found within its original image. This algorithm together with the creation of the JND threshold values is shown in algorithm 1.

Algorithm 1 Perceptual Fusion Algorithm

input : Images: S^0, S^1
$\omega^0 \leftarrow$ DT-CWT $\left(S^0\right)$
$\omega^1 \leftarrow$ DT-CWT $\left(S^1\right)$
$\omega^2 \leftarrow$ Max-Coefficient Fusion $\left(\omega^0,\omega^1\right)$ `// w^2 ≡ initial fused DT-CWT transform`
for All DT-CWT coefficients$(\forall:\ \lambda,\theta,i,j)$ **do**
  for $\forall\ I \in \{0,1,2\}$ **do**
    $v^I \equiv$ considered DT-CWT coefficient in ω^I
    $a^I_l\left(v^I\right) \leftarrow \ldots$ `// eq(4)`
    $a^I_{c\_intra}\left(v^I\right) \leftarrow \ldots$ `// eq(9)`
    $a^I_{c\_inter}\left(v^I\right) \leftarrow \ldots$ `// eq(11)`
    $a^I_c\left(v^I\right) \leftarrow a^I_{c\_intra}\left(v^I\right) a^I_{c\_inter}\left(v^I\right)$ `// eq(8)`
    $t^I_{JND}\left(v^I\right) \leftarrow a^I_l\left(v^I\right) a^I_c\left(v^I\right)$ `// eq(3)`
    $NI^I\left(v^I\right) \leftarrow \left|v^I\right| / t^I_{JND}\left(v^I\right)$ `// eq(18)`
  end
  $m \leftarrow \arg\max_I \left(NI^I,\ \forall\ I \in \{0,1\}\right)$ `// eq(20)`
  $v^3 \leftarrow v^m t^2_{JND}\left(v^2\right) / t^m_{JND}\left(v^m\right)$ `// eq(20)`
  Place v^3 into new DT-CWT transform ω^3
end
F = DT-CWT$^{-1}(\omega^3)$
output: Fused image F

C. Example Visualisation

Example visualisations of intermediate fusion measures are given in figure 5. This figure shows the initial fusion pair (top row), the JND thresholds t^I_{JND} (second row) and the fusion correction factors t^2_{JND}/t^I_{JND} (bottom row). However, as these measures are calculated on a subband by subband basis these images just show an example of a single subband (the highest frequency subband $\lambda = 1$, with orientation $\theta = 15°$).

The JND thresholds (second row) t^I_{JND} is calculated from (3). From these images it is clear that the luminance adaptation is the dominating factor. The correction factor (bottom row) shows the correction factor defined in (20). Only the

[3]This assumes that perceptual importance or "Noticeability Index" is a linear function of the JND. Perceptual importance in suprathreshold regions may in fact not be linear [42] but we make this assumption for simplicity.

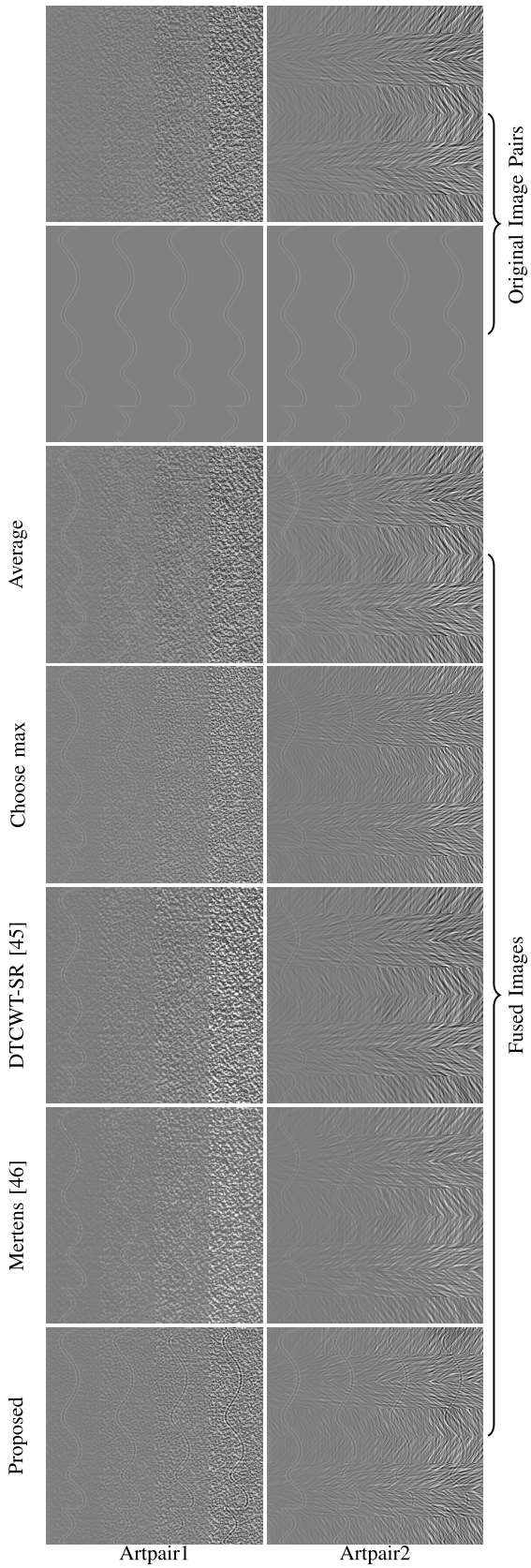

Fig. 4. Artificial image fusion.

selected coefficients are shown in each image (with unselected coefficients being zero or black). This also shows (given the JND values for each subband) that areas of high luminance

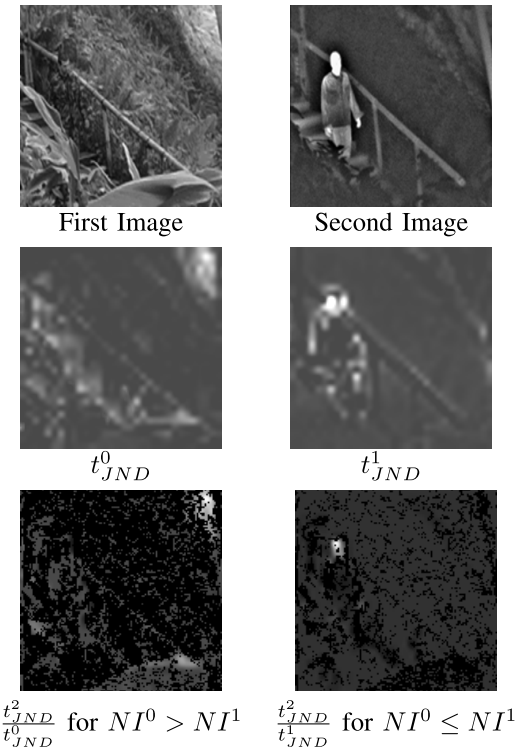

Fig. 5. Example visualisation of intermediate fusion measures. First row: Original images. Second row: JND thresholds t^I_{JND}. Bottom row: Correction factors t^2_{JND}/t^I_{JND}.

get high correction values. This row also demonstrates the stability of the correction factors for a typical subband of a typical image.

V. Results

A. Artificial Images

In order to characterise and visualise the effect of texture based contrast masking we have created two pairs of artificial images to fuse as shown in figure 4. In this figure, each column contains a distinct artificial image pair (Artpair1 and Artpair2 at the top of each column) and their associated fusion results. Both input image pairs comprise a texture of varying intensity and a discontinuity (vertical wavy lines). Both texture images were created from a single Brodatz texture (d57) [44]. This texture was chosen as it exhibits typical $1/f$ "natural image" frequency content while being homogeneous. The top left image in figure 4 shows this texture with gradually increasing contrast in vertical stripes across the entire image. The image pair shown in the left column of figure 4 is intended to illustrate isotropic intraband masking. However, in order to display and characterise the fusion output of non-isotropic texture masking we have created a second masking texture (the top-right image in figure 4). This image is also formed from 8 horizontal stripes. Each of these stripes is comprised of the original image, filtered within the frequency domain with the following filter.

$$L_n(\theta) = \begin{cases} 1 & \frac{2\pi n}{N} < \theta \leq \frac{2\pi (n+1)}{N} \\ 0 & \text{otherwise} \end{cases} \tag{21}$$

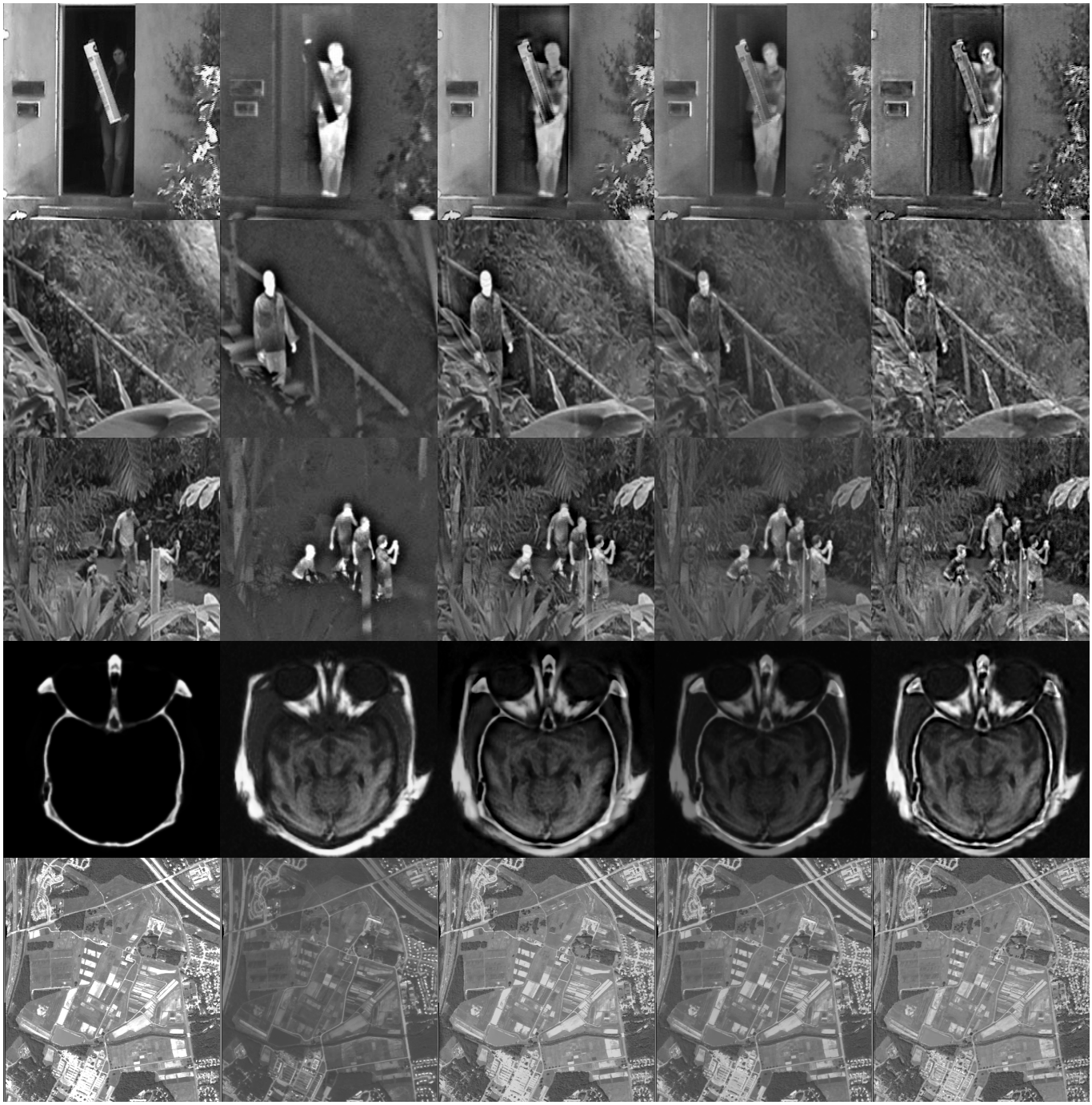

Fig. 6. Fused image results. From top to bottom the fusion pairs are labelled "Door", "Eden1", "Eden2", "Med", "Remote".

where θ is the orientation of the polar separable indices within the FFT domain, $N = 8$ (number of orientations and therefore the number of horizontal rows in the top-right figure in figure 4) and $n = \{0, \ldots, N-1\}$.

The fused output (of the proposed method) for this image pair (displayed in the same figure) shows that the magnitude of the discontinuity (vertical wavy lines) is dependent not only on the contrast of the background texture but also on its orientation (i.e. coefficients are increased in magnitude when the orientation of the chosen coefficient is similar to the masking orientation as set out in (13)).

B. Real Images

Figure 6 shows five image pairs processed using three fusion methods. The three fusion methods are:

1) DTCWT-SR: Liu *et al.* [45] proposed a combination of multiscale transforms and dictionary learned sparse representations for image fusion. Although they have compared many

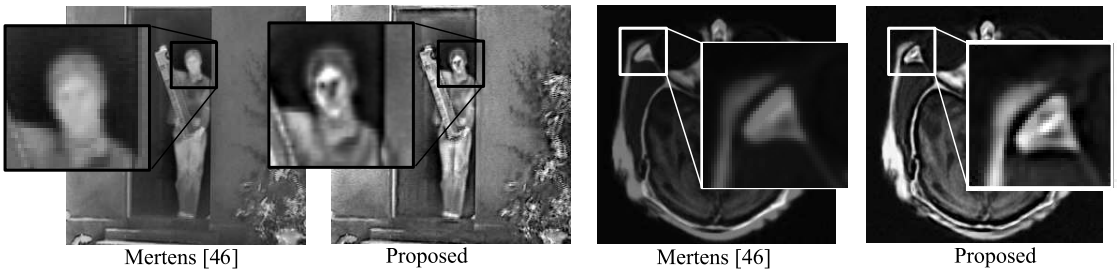

Fig. 7. Improvement of fusion results of the proposed method demonstrated through the magnification of regions within two image fusion pairs. Images are the "door" pair (left) and "med" pair (right).

different multi-scale transforms, we have used the Dual Tree Complex Wavelet Transform not only for comparability to our work, but they report it is the best performing transform within their work.

2) Mertens et al. [46]: This fusion method uses a Laplacian multi-scale decomposition that employs different types of simple perceptual fusion criteria. We have used the contrast as local criteria for fusion. This has been shown to give good results for exposure and more general types of image fusion.

3) Proposed: DT-CWT fusion used together with (3), (4) and (8) to generate the JND threshold t_{JND}. The coefficients within all the high pass subbands are then chosen and modified using (18) and (20) respectively. The Low-Low (LL) subbands are fused in the same manner as the choose-max method.

An exhaustive comparison of all possible image fusion methods was not feasible. These two comparative image fusion methods were chosen as they represent a combination of the most recent, best performing and popular methods.

The top three rows in figure 6 show registered visible and thermal IR images. These images represent typical visible/IR fusion scenarios where higher frequency visible content needs to be fused with high luminance thermal-IR areas. For important areas such as faces, conventional fusion (as exemplified by the Mertens and DTCWT-SR method) results in details that are clearly visible within the visible sources, being "washed out" (see the third and fourth column of figure 6 and the insets in figure 7).

This is due to two reasons:

- The inability of the high-pass transform coefficients of the visible image to fully capture the spatial variations within regions such as faces (due to low frequency variations of the visible image not being selected in the high luminance IR areas, such as faces).
- Luminance Masking. The luminance masking effect discussed in section III-A means that the contrast of high frequency visible content will be less perceptually important when fused within high IR luminance areas such as faces.

Adjustment of the coefficients for the proposed method increases the contrast within these regions so that they retain the same level of perceptual importance as when they were in the source image. This increase in contrast also offsets the first effect.

The fourth row of figure 6 shows a typical multimodal medical fusion application. This image pair also demonstrates how the proposed method is able to retain high frequency information from both sources within high luminance fused areas. This image pair also demonstrates texture-based contrast masking as illustrated by the skull edge (from the left image) which retains the same significant perceptual importance in the fused image as it did in the original image.

The fifth row of the same figure shows a typical remote sensing fusion application and illustrates the combination of luminance and contrast masking based perceptual fusion with the proposed method.

The first three rows of figure 6 shows the improvement of the visual results for facial regions associated with visible-thermal fusion applications. Figure 7 illustrates the improvement in fusion results in retaining visually significant content for regions that are magnified for clarity.

C. Fusion Metrics

A robust assessment of each fusion technique is essential in order to evaluate and compare the range of methods and parameters considered. Both objective and subjective assessments have been used. However, qualitative measures are time consuming to obtain and require careful control of viewing conditions and subject equivalence to render the results meaningful. Quantitative assessment techniques are therefore beneficial for effective comparisons. However, such objective measures need to be carefully aligned with subjective results. Such checks between subjective and objective measures have been done by Petrovic [47]. Where ground truth data is available, simple two-image comparison techniques such as PSNR/MSE or more perceptually aligned metrics such as SSIM can be used. For more general applications (such as medical, remote sensing and multiband fusion methods) a ground truth is not generally available and reference free quality metrics are needed. Such metrics compare the input sources together with the fused output to generate a metric quality score (ideally in the range 0.0 to 1.0) that quantifies the ability of the fusion technique to retain important visual information from each of sources. Cvejic et al. have evaluated many such metrics [48]. The relative trends of these different metrics within this paper were found to be very similar.

We have compared the results of the following fusion metrics (for the images pairs shown in figure 6): The Piella Metric Q_S [49] based on weighted SSIM values, the Xydeas metric Q_G [50] based on the image gradient,

TABLE III

RESULTS FOR FOUR FUSION METHODS AND FIVE IMAGE PAIRS. METRICS ARE Q_S: PIELLA AND HEIJMANS [49], Q_G: XYDEAS AND PETROVIC [50], Q_{MI}: HOSSNY AND NAHAVANDI [51], Q_{CV}: CHEN AND VARSHNEY [52]. LABELLING TAKEN FROM LIU *et al.* [53]

Metric	Image	Proposed	Choose Max [1]	Mertens [46]	DTCWT-SR [45]
Entropy	Door	7.5896	7.6100	**7.7179**	7.0608
	Eden1	7.2125	7.1457	**7.2421**	6.5056
	Eden2	**7.1514**	7.0925	7.1444	6.5336
	Med	6.6103	6.5315	**6.8265**	5.9540
	Remote	7.1212	7.1211	**7.1271**	6.7053
	Artpair1	**6.4145**	6.3929	5.4227	6.3963
	Artpair2	**6.3263**	6.3113	5.3438	6.3205
Q_S	Door	0.7634	**0.8448**	0.8307	0.6642
	Eden1	0.7719	**0.8471**	0.8440	0.6894
	Eden2	0.7998	**0.8582**	0.8470	0.7478
	Med	0.8014	**0.8335**	0.7639	0.5772
	Remote	0.8087	0.8540	**0.8569**	0.8259
	Artpair1	0.9708	0.9836	0.7616	**0.9850**
	Artpair2	0.9772	0.9846	0.7617	**0.9865**
Q_G	Door	0.5234	0.6554	**0.6597**	0.3511
	Eden1	0.5227	0.6582	**0.6720**	0.3432
	Eden2	0.5520	0.6651	**0.6702**	0.3791
	Med	0.6942	**0.7533**	0.6540	0.3424
	Remote	0.5513	0.6354	**0.6552**	0.5477
	Artpair1	0.8540	0.8886	0.3726	**0.8947**
	Artpair2	0.8617	0.8828	0.3836	**0.8901**
Q_{CV}	Door	**1272.2**	508.4	502.7	564.5
	Eden1	**1045.2**	203.5	181.3	376.5
	Eden2	**1266.3**	593.8	664.8	684.7
	Med	**2156.2**	1996.8	2139.3	1693.5
	Remote	**993.8**	540.9	413.1	168.8
	Artpair1	281.0	271.7	246.1	**289.0**
	Artpair2	387.9	435.9	384.0	**452.8**
Q_{MI}	Door	0.3671	**0.4788**	0.4229	0.4472
	Eden1	0.3165	**0.4661**	0.3598	0.3323
	Eden2	0.3922	0.4519	0.4431	**0.4911**
	Med	0.5975	**0.6950**	0.3515	0.6204
	Remote	0.5574	0.6149	0.6009	**0.8191**
	Artpair1	0.5121	0.5749	**0.9036**	0.7275
	Artpair2	0.5219	0.5729	0.7128	**0.8514**

the Hossny metric Q_{MI} [51] based on Mutual Information and the Chen metric Q_{CV} [52] based on a perceptual model and local image saliency. This labelling is taken from the extensive study of metric performance given by Liu *et al.* [53].

The majority of these metrics gave inferior quantitative results for our method as shown in table III. This is due to the fact that the original content has been "distorted" (i.e. amplified or attenuated) in order to create a fused output which retains the same perceptual importance as its constituent images. The exception to this was the Chen metric Q_{CV} [52] which indicated superior results for the proposed method for all the real images. This was considered to be caused by the perceptual model integrated within the Q_{CV} metric. It should be noted that the entropy of the fused output is simply a measure of information content. High values of entropy do not therefore indicate best fusion performance as fusion artefacts are included in the measure [54].

D. Perceptual Assessment

In order to assess the perceptual quality of our fusion algorithm, the use of most existing fusion metrics is not appropriate. We have therefore conducted a subjective assessment similar to that used by Petrovic [47] and Ma *et al.* [55]. Two sets of experiments were conducted, one presenting fusion results together with the original reference images and one without.

- *Reference Test:* Four images are displayed in a 2×2 grid. The original two images are displayed at the top of the screen. Two fused images using two different fusion methods are displayed at the bottom. The two displayed fusion methods are chosen from: the proposed method, the "Choose maximum" fusion method using the DT-CWT [1], the Mertens method [46] and the DTCWT-SR method [45].
- *No Reference Test:* Identical to the reference test, but the original images are not displayed.

For both tests, the two fusion images are rated by the subject on a linear scale from "good" to "bad".[4] The subjects are asked to rate the amount of information retained from both images. The worst score is 1 and the best 5 (similar to the mean opinion score scale used in compression assessment methods). The dataset used is a subset of 25 images taken from the fusion image dataset used by Petrovic [47] (comprising remote sensing, medical and visual/thermal pairs). Within the reference test, the subjects are asked to judge how much important visual information is retained from both original images, whereas within the no reference test, the subjects are asked to simply rate the quality of the fusion without seeing the originals.

1) Experimental Conditions: Seven subjects were used, two female and five male. The subjects were found to have normal or corrected to normal eyesight. The experiments were conducted in a normally lit room with a standard monitor. The viewing distance was 57cm, resulting in a viewing resolution of 29.8 pixels/degree (14.9 cycles/degree). This was used to update the value of r used in (16) and (17). All seven subjects did both the reference or non-reference evaluation with the referenced evaluation conducted first and the non-reference evaluation conducted second. Although it is possible that the subjects could learn the images and results this was considered to be a small effect due to the randomisation of the fusion methods' positions and comparisons together with the large size of the dataset.

2) Experimental Results and Discussion: As subjects do not use the same scales of quality and difference, the raw scores (labelled x_{ij} for image pair i, subject j) are normalised to Z-scores [57], [58].

$$z_{ij} = \frac{x_{ij} - \mu_x}{\sigma_x}, \tag{22}$$

where $\mu_j = \frac{1}{N}\sum_{i=1}^{N} x_{ij}$, $\sigma_j = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_{ij} - \mu_j\right)^2$ and x_{ij} $(i = 1, \ldots, N)$ are the raw scores assigned to all images by subject j. N is the number of images (in this case 25).

Figures 8 and 9 show the Z-scores of the reference and no reference experiments for each of the image pairs from 1 to 25 for the four fusion methods (proposed, choose maximum, Mertens and the DTCWT-SR methods). These figures show

[4]Task based fusion evaluation has been investigated by Dixon *et al.* [56]. However, this study used a very limited dataset. In order to evaluate our proposed method with a large range of applications (represented by the utilised dataset Petrovic [47]) such task based evaluation is not possible.

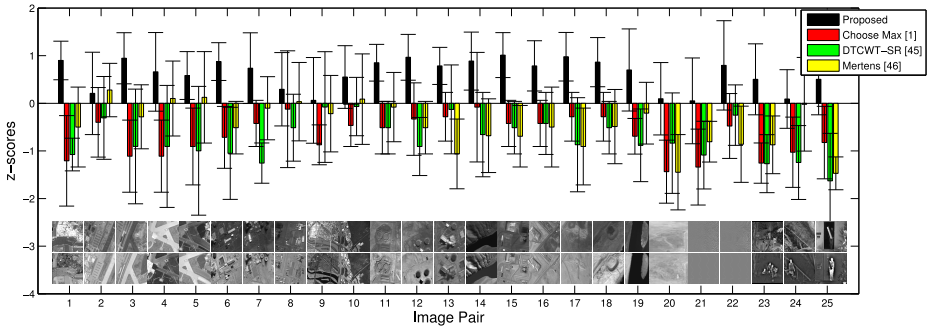

Fig. 8. Subjective tests Z-scores (no reference).

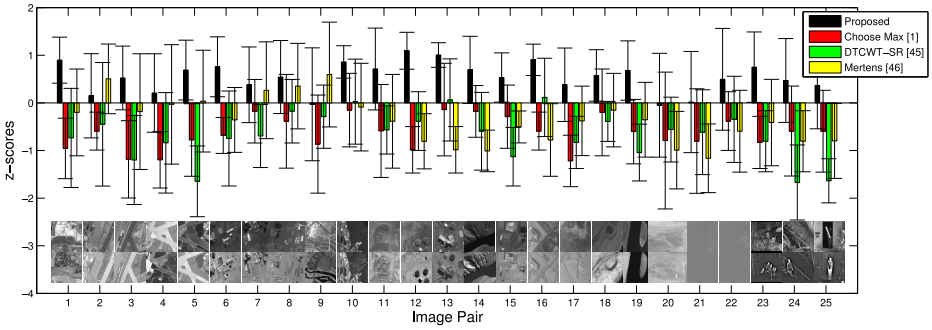

Fig. 9. Subjective tests Z-scores (fusion reference).

TABLE IV
SUBJECTIVE TEST Z-SCORE MEAN VALUES: MEAN± STANDARD ERROR

	Fusion Reference				No Reference			
Pair	Proposed	Choose Max[1]	Mertens [46]	DTCWT-SR [45]	Proposed	Choose Max[1]	Mertens [46]	DTCWT-SR [45]
1	**0.898**±0.483	-0.956±0.636	-0.735±1.040	-0.200±0.911	**0.900**±0.405	-1.208±0.950	-1.076±0.343	-0.499±0.839
2	0.149±0.884	-0.593±0.398	-0.450±1.297	**0.504**±0.732	0.208±0.865	-0.400±0.730	-0.307±0.869	**0.278**±0.560
3	**0.523**±0.669	-1.185±0.812	-1.199±0.934	-0.183±1.216	**0.945**±0.537	-1.110±0.758	-0.906±1.204	-0.284±0.672
4	**0.207**±0.824	-1.196±0.592	-0.839±1.054	-0.031±1.254	**0.660**±0.826	-1.110±0.758	-0.906±1.280	0.100±0.786
5	**0.690**±0.627	-0.779±0.762	-1.647±0.744	0.037±1.068	**0.582**±0.502	-0.907±0.806	-0.997±1.352	0.125±0.961
6	**0.758**±0.629	-0.683±0.377	-0.748±0.998	-0.358±0.681	**0.876**±0.396	-0.715±0.649	-1.051±0.966	-0.512±0.551
7	**0.380**±0.794	-0.176±0.667	-0.695±0.663	0.265±1.017	**0.738**±0.742	-0.423±0.485	-1.256±0.422	-0.103±0.662
8	**0.544**±0.766	-0.386±0.984	-0.176±0.667	0.352±0.899	**0.295**±0.770	-0.124±1.226	-0.512±0.703	0.034±0.823
9	-0.029±1.186	-0.865±1.029	-0.291±0.663	**0.595**±1.100	**0.062**±0.900	-0.872±0.419	-0.078±1.162	-0.218±0.802
10	**0.859**±0.340	-0.152±0.772	0.026±0.893	-0.089±0.921	**0.552**±0.658	-0.464±0.444	-0.068±0.612	0.087±0.950
11	**0.712**±0.859	-0.589±0.975	-0.568±0.506	-0.386±0.984	**0.852**±0.385	-0.512±0.551	-0.512±0.551	-0.080±0.727
12	**1.094**±0.388	-0.985±0.487	-0.382±0.616	-0.807±0.580	**0.966**±0.480	-0.331±0.761	-0.907±0.610	-0.512±0.551
13	**1.005**±0.258	-0.141±0.972	0.062±0.862	-0.985±0.487	**0.785**±0.389	-0.283±0.511	-0.128±0.933	-1.063±0.732
14	**0.694**±0.705	-0.179±0.543	-0.596±0.816	-1.008±0.436	**0.887**±0.609	-0.081±1.150	-0.657±0.886	-0.680±0.774
15	**0.532**±0.517	-0.291±0.663	-1.131±0.614	-0.505±0.335	**1.011**±0.472	-0.423±0.485	-0.512±0.551	-0.692±0.646
16	**0.903**±0.330	-0.593±0.398	0.115±0.822	-0.779±0.762	**0.786**±0.527	-0.423±0.485	-0.423±0.652	-0.499±0.839
17	**0.380**±0.771	-1.219±0.540	-0.830±0.547	-0.378±0.729	**0.978**±0.510	-0.283±0.511	-0.871±0.988	-0.907±0.806
18	**0.576**±0.540	-0.200±0.911	-0.390±0.420	-0.152±0.772	**0.864**±0.516	-0.283±0.511	-0.512±0.551	-0.489±0.779
19	**0.679**±0.622	-0.604±0.677	-1.044±0.596	-0.354±0.785	**0.700**±0.862	-0.692±0.374	-0.883±0.763	-0.207±0.648
20	**-0.050**±1.090	-0.790±1.436	-0.560±0.680	-0.993±0.813	**0.096**±0.760	-1.436±0.662	-0.837±1.054	-1.448±0.790
21	**0.017**±1.062	-0.802±1.098	-0.616±0.891	-1.164±0.721	**0.051**±0.901	-1.339±0.797	-1.088±0.711	-0.807±0.428
22	**0.496**±1.069	-0.382±0.616	-0.346±0.905	-0.595±0.860	**0.796**±0.938	-0.477±0.681	-0.249±0.635	-0.860±0.800
23	**0.750**±0.739	-0.830±0.547	-0.806±0.641	-0.413±0.907	**0.503**±0.745	-1.256±0.422	-1.268±0.609	-0.873±0.606
24	**0.470**±0.881	-0.591±0.945	-1.670±0.788	-0.806±0.641	**0.090**±0.616	-1.028±0.738	-1.243±0.776	-0.020±0.983
25	**0.367**±0.911	-0.595±0.860	-1.636±0.463	-0.798±0.788	**0.503**±0.745	-0.825±0.761	-1.629±0.997	-1.471±0.345

the mean and standard deviation of the Z-scores for each image pair. Table IV shows these results in tabular form and clearly shows that the $Z-$score for each image pair is higher for the proposed method in all but three cases (two for the reference case and one for the no-reference case). The results show that the proposed method is considered to be of higher quality for both sets of tests. To quantify this, a hypothesis is defined as: the mean score of the proposed method is higher than the mean of the choose-maximum method for one of the experiment types.

In order to test this hypothesis, a right-tailed, unpaired, t-test was done on the data for each image. The resulting p-value was calculated for each image pair (the probability that the null hypothesis was true). The average p-value for all images was 0.0155. This analysis was repeated for the Mertens method and the DTCWT-SR method resulting in an average p-values of 0.0796 and 0.0294 respectively. It can therefore be concluded that the average normalised score for the proposed method is greater than the three alternative methods (choose max, Mertens and DTCWT-SR) with a confidence of over 92%.

VI. Conclusion

Conventional transform-based image fusion algorithms implicitly assume that there is a simple linear relationship between coefficient magnitude and perceptual importance. This is a gross simplification. Our work has addressed this by producing a principled model of the perceptual importance of coefficients within image fusion and evaluating its performance objectively and subjectively across a representative dataset.

The results clearly show qualitative improvements in image fusion results where regions that are saturated for conventional methods now retain important perceptual content from both input images. Quantitative improvements of information content over comparable techniques are also demonstrated.

The proposed method has therefore been demonstrated to form a fused output that not only contains the most perceptually important content from the input images but is able to present the retained information with its original perceptual importance.

Subjective test results show that the proposed method can be considered the best perceptually performing method compared to the other considered fusion algorithms with a high confidence of over 92%.

It should be noted however, that due to the extremely large number of possible parametric and algorithmic variations, the results of the subjective tests can only be used to gain a sense of direction in the research.

References

[1] P. R. Hill, C. N. Canagarajah, and D. R. Bull, "Image fusion using complex wavelets," in *Proc. 13th Brit. Mach. Vis. Conf.*, 2002, pp. 487–496.

[2] H. Li, B. S. Manjunath, and S. K. Mitra, "Multi-sensor image fusion using the wavelet transform," in *Proc. IEEE Int. Conf. Image Process.*, vol. 1. Nov. 1994, pp. 51–55.

[3] N. Kingsbury, "The dual-tree complex wavelet transform: A new technique for shift invariance and directional filters," in *Proc. IEEE Digit. Signal Process. Workshop*, Aug. 1998, pp. 319–322.

[4] V. S. Petrovic and C. S. Xydeas, "Gradient-based multiresolution image fusion," *IEEE Trans. Image Process.*, vol. 13, no. 2, pp. 228–237, Feb. 2004.

[5] O. Rockinger, "Pixel-level fusion of image sequences using wavelet frames," in *Proc. 16th Leeds Appl. Shape Res. Workshop*, Jul. 1996, pp. 149–154.

[6] A. Toet, "Hierarchical image fusion," *Mach. Vis. Appl.*, vol. 3, no. 1, pp. 3–11, 1990.

[7] J. J. Lewis, R. J. O'Callaghan, S. G. Nikolov, D. R. Bull, and C. N. Canagarajah, "Pixel and region-based image fusion with complex wavelets," *Inf. Fusion*, vol. 8, no. 2, pp. 119–130, 2007.

[8] S. Nercessian, K. Panetta, and S. Agaian, "Human visual system-based image fusion for surveillance applications," in *Proc. IEEE Int. Conf. Syst., Man Cybern.*, Oct. 2011, pp. 2687–2691.

[9] S. Li, J. T. Kwok, and Y. Wang, "Multifocus image fusion using artificial neural networks," *Pattern Recognit. Lett.*, vol. 23, no. 8, pp. 985–997, 2002.

[10] J. W. G. Bhatnagar and Z. Liu, "Human visual system inspired multi-modal medical image fusion framework," *Expert Syst. Appl.*, vol. 40, no. 5, pp. 1708–1720, 2013.

[11] M. Li, W. Cai, and Z. Tan, "A region-based multi-sensor image fusion scheme using pulse-coupled neural network," *Pattern Recognit. Lett.*, vol. 27, no. 16, pp. 1948–1956, 2006.

[12] J. Huang, Y. Shi, and X. Dai, "A segmentation-based image coding algorithm using the features of human vision system," *J. Image Graph.*, vol. 4, no. 5, pp. 400–404, 1999.

[13] C. Wang and Z.-F. Ye, "Perceptual contrast-based image fusion: A variational approach," *Acta Autom. Sinica*, vol. 33, no. 2, pp. 132–137, 2007.

[14] J. A. Ferwerda, S. N. Pattanaik, P. Shirley, and D. P. Greenberg, "A model of visual masking for computer graphics," in *Proc. SIGGRAPH*, 1997, pp. 173–182.

[15] S. J. Daly, "Visible differences predictor: An algorithm for the assessment of image fidelity," *Digit. Images Human Vis.*, vol. 1666, pp. 179–206, Aug. 1993.

[16] P. G. Barten, *Contrast Sensitivity of the Human Eye and Its Effects on Image Qualityy*. Bellingham, WA, USA: SPIE, 1999.

[17] A. B. Watson and A. J. Ahumada, "A standard model for foveal detection of spatial contrast," *J. Vis.*, vol. 5, no. 9, pp. 717–740, 2005.

[18] R. Mantiuk, K. J. Kim, A. G. Rempel, and W. Heidrich, "HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions," *ACM Trans. Graph.*, vol. 30, no. 4, pp. 40:1–40:14, 2011.

[19] I. Höntsch and L. J. Karam, "Adaptive image coding with perceptual distortion control," *IEEE Trans. Image Process.*, vol. 11, no. 3, pp. 213–222, Mar. 2002.

[20] Z. Liu, L. J. Karam, and A. B. Watson, "JPEG2000 encoding with perceptual distortion control," *IEEE Trans. Image Process.*, vol. 15, no. 7, pp. 1763–1778, Jul. 2006.

[21] A. J. Ahumada, Jr., and H. A. Peterson, "Luminance-model-based DCT quantization for color image compression," *Proc. SPIE*, vol. 1666, p. 365, Dec. 1992.

[22] A. B. Watson, "DCTune: A technique for visual optimization of DCT quantization matrices for individual images," in *SID Int. Symp. Dig. Tech. Papers*, vol. 24. 1993, pp. 946–949.

[23] K.-K. Ma and L. Huang, "Perceptually based subband AMBTC image coder," in *Proc. ICICS, Int. Conf. Inf., Commun. Signal Process.*, vol. 1. Sep. 1997, pp. 9–12.

[24] M. Nadenau, "Integration of human color vision models into high quality image compression," Ph.D. dissertation, Faculté Sci. Techn. l'ingénieur, École Polytechn. Fédérale Lausanne, Lausanne, Switzerland, 2000.

[25] X. Zhang, W. Lin, and P. Xue, "Improved estimation for just-noticeable visual distortion," *Signal Process.*, vol. 85, no. 4, pp. 795–808, Apr. 2005.

[26] Z. Wei and K. N. Ngan, "Spatio-temporal just noticeable distortion profile for grey scale image/video in DCT domain," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 3, pp. 337–346, Mar. 2009.

[27] X. Zhang, W. Lin, and P. Xue, "Just-noticeable difference estimation with pixels in images," *J. Vis. Commun. Image Represent.*, vol. 19, no. 1, pp. 30–41, 2008.

[28] C.-H. Chou and Y.-C. Li, "A perceptually tuned subband image coder based on the measure of just-noticeable-distortion profile," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 5, no. 6, pp. 467–476, Dec. 1995.

[29] C.-H. Lee, P.-Y. Lin, L.-H. Chen, and W.-K. Wang, "Image enhancement approach using the just-noticeable-difference model of the human visual system," *J. Electron. Imag.*, vol. 21, p. 033007, Jul. 2012.

[30] X. Yang *et al.*, "Rate control for videophone using local perceptual cues," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 4, pp. 496–506, Apr. 2005.

[31] R. Safranek and J. Johnston, "A perceptually tuned sub-band image coder with image dependent quantization and post-quantization data compression," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 1989, pp. 1945–1948.

[32] M. Miloslavski and Y.-S. Ho, "Zerotree wavelet image coding based on the human visual system model," in *Proc. IEEE Asia–Pacific Conf. Circuits Syst., Microelectron. Integr. Syst. (APCCAS)*, Nov. 1998, pp. 57–60.

[33] Y. Zhang, M. Naccari, D. Agrafiotis, M. Mrak, and D. R. Bull, "High dynamic range video compression exploiting luminance masking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 5, pp. 950–964, May 2016.

[34] J. M. Foley, "Human luminance pattern-vision mechanisms: Masking experiments require a new model," *J. Opt. Soc. Amer. A, Opt., Image Sci., Vis.*, vol. 11, no. 6, pp. 1710–1719, 1994.
[35] J. Foley and G. Boynton, "A new model of human luminance pattern vision mechanisms: Analysis of the effects of pattern orientation, spatial phase and temporal frequency," *Proc. SPIE*, vol. 2054, pp. 32–42, Mar. 1994.
[36] G. E. Legge and J. M. Foley, "Contrast masking in human vision," *J. Opt. Soc. Amer.*, vol. 70, no. 12, pp. 1458–1471, Dec. 1980.
[37] P. Hill, A. Achim, M. E. Al-Mualla, and D. Bull, "Contrast sensitivity of the wavelet, dual tree complex wavelet, curvelet, and steerable pyramid transforms," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2739–2751, Jun. 2016.
[38] A. Watson, "DCT quantization matrices visually optimized for individual images," *Proc. SPIE*, vol. 1913, pp. 202–216, Sep. 1993.
[39] A. B. Watson and J. Solomon, "Model of visual contrast gain control and pattern masking," *J. Opt. Soc. Amer. A, Opt., Image Sci., Vis.*, vol. 14, no. 9, pp. 2379–2391, 1997.
[40] D. Bull, *Communicating Pictures*. San Diego, CA, USA: Academic, 2014.
[41] A. B. Watson, G. Y. Yang, J. A. Solomon, and J. Villasenor, "Visibility of wavelet quantization noise," *IEEE Trans. Image Process.*, vol. 6, no. 8, pp. 1164–1175, Aug. 1997.
[42] D. M. Chandler and S. S. Hemami, "Additivity models for suprathreshold distortion in quantized wavelet-coded images," *Proc. SPIE*, vol. 4662, pp. 105–117, Jan. 2002.
[43] A. Loza, A. Achim, D. Bull, and N. Canagarajah, "Statistical image fusion with generalised Gaussian and alpha-stable distributions," in *Proc. 15th Int. Conf. Digit. Signal Process.*, Jul. 2007, pp. 268–271.
[44] P. Brodatz, *Textures: A Photographic Album for Artists and Designers*. New York, NY, USA: Dover, 1966.
[45] Y. Liu, S. Liu, and Z. Wang, "A general framework for image fusion based on multi-scale transform and sparse representation," *Inf. Fusion*, vol. 24, pp. 147–164, Jul. 2015.
[46] T. Mertens, J. Kautz, and F. V. Reeth, "Exposure fusion," in *Proc. 15th Pacific Conf. Comput. Graph. Appl.*, Oct. 2007, pp. 382–390.
[47] V. Petrovic, "Subjective tests for image fusion evaluation and objective metric validation," *Inf. Fusion*, vol. 8, no. 2, pp. 208–216, 2007.
[48] N. Cvejic, J. Lewis, D. R. Bull, and C. N. Canagarajah, "Adaptive region-based multimodal image fusion using ICA bases," *Inf. Fusion*, pp. 288–293, Jul. 2006.
[49] G. Piella and H. Heijmans, "A new quality metric for image fusion," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2003, pp. 176–203.
[50] C. Xydeas and V. Petrović, "Objective image fusion performance measure," *Electron. Lett.*, vol. 36, no. 4, pp. 308–309, 2000.
[51] M. Hossny and S. Nahavandi, "Image fusion algorithms and metrics duality index," in *Proc. IEEE Image Process. Int. Conf. (ICIP)*, Jan. 2009, pp. 2193–2196.
[52] H. Chen and P. K. Varshney, "A human perception inspired quality metric for image fusion based on regional information," *Inf. Fusion*, vol. 8, no. 2, pp. 193–207, 2007.
[53] Z. Liu, E. Blasch, Z. Xue, J. Zhao, R. Laganiere, and W. Wu, "Objective assessment of multiresolution image fusion algorithms for context enhancement in night vision: A comparative study," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 94–109, Jan. 2012.
[54] D. C. M. Hossny, S. Nahavandi, and A. Bhatti, "Towards autonomous image fusion," in *Proc. IEEE Int. Conf. Control Autom. Robot. Vis. (ICARCV)*, Dec. 2010, pp. 1748–1754.
[55] K. Ma, K. Zeng, and Z. Wang, "Perceptual quality assessment for multi-exposure image fusion," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3345–3356, Nov. 2015.
[56] T. D. Dixon *et al.*, "Methods for the assessment of fused images," *ACM Trans. Appl. Perception*, vol. 3, no. 3, pp. 309–332, 2006.
[57] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1427–1441, Jun. 2010.
[58] K. Seshadrinathan, H. R. Sheikh, Z. Wang, and A. C. Bovik, "Structural and information theoretic approaches to image quality assessment," in *Image Fusion and Its Applications*, Z. Liu and R. Blum, Eds. Boca Raton, FL, USA: CRC Press, 2005, pp. 473–499.

Paul Hill (M'12) received the B.Sc. degree from the Open University in 1996, and the M.Sc. degree and the Ph.D. degree in image segmentation and fusion from the University of Bristol, Bristol, U.K., in 1998 and 2002, respectively. He is currently a Senior Research Fellow with the Department of Electrical and Electronic Engineering, University of Bristol, and also lecturing in audio technology. His research interests include image and video analysis, compression, and fusion.

Mohammed Ebrahim Al-Mualla received the B.Eng. degree in telecommunications from the Etisalat College of Engineering, United Arab Emirates, and the M.Sc. degree in communication systems and signal processing and the Ph.D. degree in electrical and electronics engineering from the University of Bristol, U.K. He is currently the Senior Vice President of Research and Interim Provost, Kahlifa University, United Arab Emirates.

David Bull (M'94–SM'07–F'12) received the B.Sc. degree from the University of Exeter in 1980, the M.Sc. degree from the University of Manchester in 1983, and the Ph.D. degree from the University of Cardiff in 1988. He was the Head of the Electrical and Electronic Engineering Department, University of Bristol, from 2001 to 2006. He is currently the Head of the Visual Information Laboratory and also the Director of the Bristol Vision Institute. He has authored over 400 academic papers. His current activities are focused on the problems of image and video communications and analysis for wireless, Internet, military, and broadcast applications. He currently holds the Chair in signal processing with the University of Bristol.