A Nonreference Image Fusion Metric Based on the Regional Importance Measure

Nedeljko Cvejic, *Member, IEEE*, Tapio Seppänen, and Simon J. Godsill, *Member, IEEE*

***Abstract*—In this paper, we present a novel metric for evaluation of image fusion algorithms, based on evaluation of similarity of regions in images to be fused with the corresponding regions in the fused image. The metric uses several factors to quantify the importance of regions in each of the input images, such as contrast, size, and shape of region. The similarity of the corresponding regions in an input image and the fused image is measured using a wavelet-based mutual information measure. Experimental results show that the proposed metric's ranking of different image fusion methods is more consistent with the subjective quality of the fused image than the state-of-the-art image fusion metrics.**

***Index Terms*—Image fusion, image fusion metric, region-based similarity, Visual Information Fidelity.**

I. Introduction

IMAGE and video fusion is emerging as a vital technology in many military, surveillance and medical applications. It is a subarea of the more general topic of data fusion, dealing with image and video data [1], [2]. The ability to combine complementary information from a range of distributed sensors with different modalities can be used to provide enhanced performance for visualization, detection or classification tasks. Multisensor data often present complementary information about the scene or object of interest, and thus image fusion provides an effective method for comparison and analysis of such data. There are several benefits of multisensor image fusion: wider spatial and temporal coverage, extended range of operation, decreased uncertainty, improved reliability, and increased robustness of the system performance.

In several application scenarios, image fusion is only an introductory stage to another task, e.g., human monitoring. Therefore, the performance of the fusion algorithm must be measured in terms of improvement in the subsequent tasks. For example, in classification systems, the common evaluation measure for image fusion algorithms is the number of the correct classifications in the fused image. In the surveillance application scenario, a human operator using a suitably fused representation of infrared (IR) and visual imagery is able to construct a more complete mental representation of the perceived scene, resulting in an increased situational awareness.

In many applications, the human perception of the fused image is of fundamental importance and as a result the fusion results are mostly evaluated by subjective criteria [3], [4]. Objective image fusion performance evaluation is a tedious task due to different application requirements and the lack of a clearly defined ground-truth. Various fusion algorithms presented in the literature [1], [5] have been evaluated objectively by constructing an "ideal" fused image and using it as a reference for comparison with the experimental results [6], [7]. Mean squared error (MSE)-based metrics were widely used for these comparisons. Several objective performance metrics for image fusion have been proposed where the knowledge of ground-truth is not assumed. An overview of the state-of-the-art image fusion metrics, which are commonly used for evaluation of fusion algorithm is given in Section II.

In this paper, we present a novel objective nonreference quality assessment metric for image fusion. It takes into account region-based measurements to estimate how well the important information in the source images is represented by the fused image. The rest of the paper is organized as follows. Section II gives an overview of the most important existing image fusion metrics. Section III presents the detailed description of the proposed metric, including the calculation of the importance of the regions in the input images, whereas Section IV contains experimental results obtained by using the proposed metric and a comparison with the state-of-the-art methods.

II. Overview of Fusion Metrics

The existing metrics for evaluation of image fusion algorithms are generally based on a measurement of the fidelity of the transfer of a feature (e.g., edges, amount of information) from the input images to the fused output. Most of objective performance measures for image fusion are not based on the use of the ground-truth data to evaluate a fusion algorithm, i.e., the performance is calculated using two input images and the fused output [30], [31]. Apart from using them to evaluate image fusion processes, the metrics were also used to optimize image fusion algorithms to produce the highest visual quality of the fused image [8], [9].

A. Mutual Information Metric

Mutual information metric has emerged as an alternative for measuring image fusion performance. It is based on the measure

Manuscript received April 30, 2008; revised December 04, 2008. Current version published March 11, 2009. This work was supported by the U.K. Data and Information Fusion Defence Technology Center (DIF DTC). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Zhou Wang.

N. Cvejic and S. J. Godsill are with the Signal Processing and Communications Laboratory, Department of Engineering, University of Cambridge, Cambridge, CB2 1PZ, U.K. (e-mail: nc332@cam.ac.uk).

T. Seppänen is with the Computer Engineering Laboratory, Department of Electrical and Information Engineering, University of Oulu, FIN-90014 Oulu, Finland.

Digital Object Identifier 10.1109/JSTSP.2009.2015071

of the degree of dependence of the two random variables A and B. It is defined by Kullback–Leibler measure [13]

$$I_{AB}(a,b) = \sum_{x,y} p_{AB}(a,b) \cdot \log \frac{p_{AB}(a,b)}{p_A(a) \cdot p_B(b)} \quad (1)$$

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

$$I_{FA}(f,a) = \sum_{x,y} p_{FA}(f,a) \cdot \log \frac{p_{FA}(f,a)}{p_F(f) \cdot p_A(a)} \quad (2)$$

$$I_{FB}(f,b) = \sum_{x,y} p_{FB}(f,b) \cdot \log \frac{p_{FB}(f,b)}{p_F(f) \cdot p_B(b)} \quad (3)$$

and the image fusion performance measure can be defined as [13]

$$M_F^{AB} = I_{FA}(f,a) + I_{FB}(f,b). \quad (4)$$

In addition to the basic mutual information metric, another metric has been proposed, which uses the MI concept, but uses Tsallis entropy to calculate the degree of dependence between the input pixels and the pixels in the fused image [14].

B. Piella Metric

The measure used as the basis for the Piella metric is the Universal Image Quality Index (UIQI). The authors compared the proposed quality index to the standard MSE quality measure and concluded that the new index outperforms the MSE, due to the UIQI's ability to measure structural distortions [10].

Let $X = x_i | i = 1, 2, \ldots, N$ and $Y = y_i | i = 1, 2, \ldots, N$ be the original and the test image signals, respectively. The proposed quality index was defined by Wang and Bovik [10] as

$$Q = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \frac{2\overline{x}\,\overline{y}}{(\overline{x})^2 + (\overline{y})^2} \frac{2\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2} \quad (5)$$

where

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i, \qquad \overline{y} = \frac{1}{N} \sum_{i=1}^{N} y_i \quad (6)$$

$$\sigma_x^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x})^2, \sigma_y^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \overline{y})^2 \quad (7)$$

$$\sigma_{xy} = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x})(y_i - \overline{y}). \quad (8)$$

When written as above, UIQI models image distortions as a product of three components: loss of correlation, luminance distortion, and contrast distortion. The dynamic range of Q is $[-1, 1]$. The best value 1 is achieved if and only if $y_i = x_i$ for all $i = 1, 2, \ldots, N$. Since images are generally nonstationary signals, it is appropriate to measure Q_0 over local regions and then combine the different results into a single measure Q. In [10], the authors propose a sliding window: starting from the top-left corner of the two images X, Y, a sliding window of a fixed size traverses over the entire image until the bottom-right corner is reached. For each window w the local quality index $Q_0(X, Y|w)$ is computed. Finally, the overall image quality index Q is computed by averaging all local quality indices. Wang and Bovik have compared (under several types of distortions) their quality index with existing image measures such as MSE as well as with subjective evaluations. The main finding was that UIQI outperforms the MSE, which due to the index's ability of measuring structural distortions, in contrast to the MSE which is highly sensitive to the energy of errors.

In order to apply the UIQI for image fusion evaluation, Piella and Heijmans [11] introduce salient information to the metric

$$Q_p(X,Y,F) = \sum_{w \in W} c(w)[\lambda Q(X,F|w) + (1-\lambda) Q(Y,F|w)] \quad (9)$$

where X and Y are the input images, F is the fused image, $c(w)$ is the overall saliency of a window, and λ is defined as

$$\lambda = \frac{s(X|w)}{s(X|w) + s(Y|w)} \quad (10)$$

should reflect the relative importance of image X compared to image Y within the window w. $s(X|w)$ denotes saliency of image X in window w. Finally, to take into account aspects of the human visual system (HVS), the same measure is computed with "edge images" (X', Y' and F') instead of the grayscale images X, Y and F

$$Q_E(X,Y,F) = Q_p(X,Y,F)^{1-\alpha} Q_p(X',Y',F')^{\alpha}. \quad (11)$$

C. Petrovic Metric

The fusion metric proposed Petrovic and Xydeas [12] is obtained by evaluating the relative amount of edge information transferred from the input images to the output image. It also takes into account the relative perceptual importance of the visual information found in the input images, by assigning perceptual importance weights to more salient edges. 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 $A_{AF}(n,m)$, respectively, of an input image A with respect to the fused one F are defined as [12]

$$G^{AF}(n,m) = \begin{cases} \frac{g_F(n,m)}{g_A(n,m)}, & \text{if } g_A(n,m) > g_F(n,m) \\ \frac{g_A(n,m)}{g_F(n,m)}, & \text{otherwise} \end{cases} \quad (12)$$

$$A^{AF}(n,m) = \frac{\left|\left|\alpha_A(n,m) - \alpha_F(n,m)\right| - \frac{\pi}{2}\right|}{\frac{\pi}{2}}. \quad (13)$$

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

$$Q_g^{AF}(n,m) = \frac{\Gamma_g}{1 + e^{k_g(G^{AF}(n,m) - \sigma_g)}} \quad (14)$$

$$Q_\alpha^{AF}(n,m) = \frac{\Gamma_\alpha}{1 + e^{k_\alpha(A^{AF}(n,m) - \sigma_\alpha)}} \quad (15)$$

where the constants Γ_g, k_g, σ_g and Γ_α, k_α, σ_α 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

$$Q^{AF}(n,m) = Q_g^{AF}(n,m) \cdot Q_\alpha^{AF}(n,m), 0 \le Q^{AF}(n,m) \le 1. \tag{16}$$

Having $Q^{AF}(n,m)$ and $Q^{BF}(n,m)$ a normalized weighted performance metric of a given process p that fuses A and B into F is given as

$$Q_p = \frac{\sum_{n=1}^{N}\sum_{m=1}^{M} Q^{AF}(n,m)w_A(n,m) + Q^{BF}(n,m)w_B(n,m)}{\sum_{n=1}^{N}\sum_{m=1}^{M} w_A(n,m) + w_B(n,m)}. \tag{17}$$

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.

III. Proposed Method for Objective Image Fusion Evaluation

One of the main shortcomings of the existing methods for evaluation of performance of image fusion algorithms is lack of estimation of the extent to which the important objects (regions in the images to be fused) should be transferred to the fused image. For example, in a multimodal image fusion scenario, it is very common to have only one or a few objects in the IR image that are of high importance, such as a person walking, whereas the rest of the image is considerably less informative. If an image fusion algorithm fails to incorporate the most important object(s) from IR image to the fused image, the usefulness of the fused image is significantly decreased, regardless of how the remaining details from both input images are transferred to the fused one.

Therefore, we adopt a novel approach to objective image fusion evaluation, which quantifies the importance of each of the input regions in the images to be fused and then measures how faithfully the given region is transferred into the fused image. Consequently, one of the essential duties of the proposed image fusion metric is to asses precisely the way the HVS determines the importance of the regions in an observed scene.

A. Image Regions Important for the Human Visual System

In order to efficiently process the high volumes of information present in a visual field, the HVS operates using an adaptive resolution. Although our field of view is roughly 180° horizontally and 140° vertically, a high degree of visual acuity by HVS is obtained over a very small area ($\sim$2° in diameter) called the fovea. Therefore, an accurate inspection of the scattered objects in our visual field requires eye movements. Rapid shifts in the eye's focus of attention (saccades) occur every 100–500 ms. Visual attention mechanisms are used to control these saccades. The pre-attentive vision of HVS operates in parallel, looking in the periphery for important areas and uncertain areas for the eye to focus on at the next saccade [15]. Thus, a very strong relationship exists between eye movements and attention.

Existing studies on human eye movement for both images and video indicate that eye movements are indeed highly correlated amongst subjects. Results in [16] demonstrated that a strong correlation between viewer eye movements exists, given that the subjects were observing a given image in the same context (i.e., with the same instructions and motivation). It was also demonstrated that even if given unlimited viewing time, a human observer will attend only to a handful of important regions which continually attract her/his attention [16]. Similar results have been confirmed in video sequences [17]. Thus, all these experiment suggests that eye movements are not individual, and that there is a strong correlation between the direction of gaze of different subjects, if looking at an image in the same context.

In order to determine the importance of the different regions in an image, we need to evaluate the factors which affect visual attention. Human visual attention is dominated by high and low level factors. High level factors usually involve some feedback process from memory and may involve template matching. Low level processes are generally fast, feed-forward mechanisms involving relatively simple processing. In general, the objects that stand out from their immediate surroundings are more likely to attract our attention, since one of the main goals of the HVS is to minimize uncertainty. Low level factors which have been found to influence visual attention include:

1) Size: Results in [18] have shown that region size has an important effect in attracting attention of the HVS. Larger regions are more likely to draw our attention than smaller ones, however only up to a saturation point, after which the importance due to region size decreases.

2) Color: Color has been found to be important in attracting attention [19]. A number of colors (such as red) have been shown to draw our attention and a strong color impact occurs when the color of a region is distinct from the color of its background.

3) Contrast: The HVS converts luminance into contrast at an early stage of processing. Region contrast is therefore a very strong low-level visual attractor. Regions which have a high contrast with their surroundings attract our attention and are likely to be of greater visual importance [18].

4) Shape: Elongated and thin regions (edge-like) have been found to be visual attractors [20]. They are more likely to attract visual attention than circular regions of the same area and contrast.

A number of additional low level factors which have been found to influence attention include brightness and orientation. The present literature also lists several high level factors, some of which are as follows.

5) Location: Eye-tracking experiments have shown that viewers eyes are directed at the center 25% of a screen for a majority of visual experiments [21].

6) Humans: Many studies have demonstrated that humans are drawn to focus on people in a scene, in particular their faces, eyes, mouth, and hands [20].

7) Context: Viewers' eye movements can be dramatically changed, depending on the instructions they are given while observing the image [20].

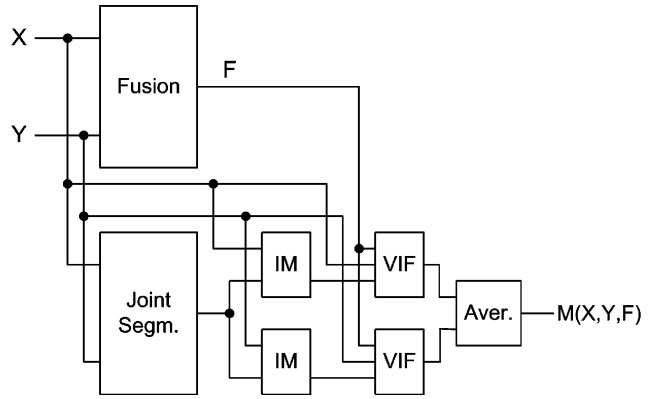

Fig. 1. Overall scheme of the proposed image fusion metric. X and Y are input images, F represents the fused image. IM stands for importance calculation for each of the regions determined by joint segmentation. VIF block calculates the similarity of the regions between X and F and between Y and F, weighted by the region importance information. The final value is calculated as the mean value of the values obtained for each of the images to be fused.

Even though the aforementioned factors that influence visual attention have been identified, there has been limited research on the exact weighting of the given factors and their dependencies. A number of factors are clearly of very high importance, but it is difficult to determine exactly the exact hierarchy of importance. A particular factor may be more important than another factor in one image, while in another image the opposite may be the case. Due to this lack of information, it is necessary to consider a large number of factors when modeling visual attention [22]. It is also desirable that the factors used be independent, so that a particular factor would not exhibit excessive influence on the overall importance.

The overall scheme of the proposed image fusion metric is given in Fig. 1. Input images are first segmented into regions using joint segmentation and then the visual importance of each region calculated. Finally, the similarity between the identified regions is measured and the metric value calculated by weighting each regional similarity by the importance of the given region. In the following subsections we give an overview of each of the steps in the calculating the final metric value.

B. Extracting Regions Using Joint Segmentation

The quality of the segmentation algorithm is of the vital importance to the accuracy of the metric. An adapted version of the combined morphological–spectral unsupervised image segmentation algorithm is used, which is described in [23], enabling it to handle multimodality images.

The segmentation algorithm works in two stages. The first stage produces an initial segmentation by using both textured and nontextured queues. The detail coefficients of the DT-CWT are used to process texture. The complex magnitude of the coefficients of each DT-CWT subband can be used to characterize the local texture content. This is because the basis functions from each DT-CWT subband very closely resemble Gabor filters. The results of processing of textured queues are utilized as input to the following step in the segmentation process, which is to derive the perceptual gradient. The gradient function is applied to all levels and orientations of the DT–CWT coefficients and up-sampled to be combined with the gradient of the intensity information to give a perceptual gradient. The larger gradients indicate possible edge locations. The watershed transform [32] of the perceptual gradient gives an initial segmentation. The second stage uses these primitive regions to produce a graph representation of the image which is processed using a spectral clustering technique [23].

Regions with a size less than 16 pixels are merged with their most similar neighbor, to avoid excessively small regions. Merging of region of size smaller than 16 pixels was an experimental criterion, based on extensive tests performed on a large dataset of fused images. This value for the smallest allowed region was kept the same throughout the examples presented in the paper. However, this value can be decreased in applications that deal with very small regions, such as hyperspectral image fusion/processing.

The segmentation can be performed either separately or jointly. For separate segmentation, each of the input images generates an independent segmentation map for each image

$$S_1 = \sigma(i_1, D_1), \ldots, S_N = \sigma(i_N, D_N) \tag{18}$$

where σ represents segmentation process, and D_n represent detail coefficients of the dual-tree complex wavelet transform (DT-CWT) used in segmentation of image i_n. Alternatively, information from all images could be used to produce a joint segmentation map.

$$S_{\text{joint}} = \sigma(i_1 \cdots i_N, D_1 \cdots D_N). \tag{19}$$

In general, jointly segmented images work better for image fusion applications [2]. This is because the segmentation map will contain a minimum number of regions to represent all the features in the scene most efficiently. A problem can occur for separately segmented images, where different images have different features or features which appear as slightly different sizes in different modalities. Where regions partially overlap, if the overlapped region is incorrectly dealt with, artifacts will be introduced and the extra regions created to deal with the overlap will increase the time taken to fuse the images. After the regions are determined by joint segmentation, the importance of each of the regions in the images to be fused is calculated (Fig. 1).

C. Quantification of Importance of an Image Region

The segmented image is analyzed by a number of different factors known to influence attention [24], and an importance is assigned to each region for each factor. We have chosen five different importance factors in this algorithm, although the flexible structure of this approach allows additional factors to easily be incorporated. The different factors that we have chosen are as follows.

1) Contrast of Region: The contrast importance I_1 of a region R_i is calculated as

$$I_1(R_i) = \overline{R}_i - \overline{R}_{i-nb} \tag{20}$$

where $\overline{R}_i$ is the mean value of the pixel values in the region R_i, and $\overline{R}_{i-nb}$ is the mean grey-level of all of the neighboring

regions of R_i. Subtraction is used rather than division, since it is assumed that the grey-levels are a perceptually linear space. I_1 is scaled to the range [0, 1] using the minimum and maximum value of $I_1(R_i)$ in the analyzed image.

2) Size of Region: Importance due to region size is calculated as

$$I_2(R_i) = \max\left(\frac{A(R_i)}{A_{\max}}, 1\right) \tag{21}$$

where $A(R_i)$ is the area of region R_i in pixels, and is $A_{\max}$ a constant used to prevent excessive weighting being given to very large regions. We have set $A_{\max}$ to be equal to 3% of the total image area.

3) Location of Region: Importance due to location of a region is calculated as

$$I_3(R_i) = \frac{\text{center}(R_i)}{A(R_i)} \tag{22}$$

where $\text{center}(R_i)$ is the number of pixels in region R_i which are also in the center 25% of the image. Thus, regions contained entirely within the central quarter of an image will have a location importance of 1, and regions with no central pixels will have $I_3(R_i) = 0$.

4) Shape of Region: Importance due to region shape is calculated as

$$I_4(R_i) = \frac{bp(R_i)^s}{A(R_i)} \tag{23}$$

where $bp(R_i)$ is the number of pixels in the region R_i which border with other regions, and s is a constant. Experiments showed that a value of s of 1.75 provides a good discrimination of shapes. This importance factor will enable elongated regions to have a high shape importance, while it will decrease the importance of the rounder regions. I_4 is scaled to the range [0, 1] using the maximum value of $I_4(R_i)$ in the analyzed image.

5) Region in Foreground/Background: We detect background regions by determining the proportion of the total image border that is contained in each region. Regions with a high number of image border pixels will be classified as belonging to the background and will have a low foreground importance as given by

$$I_5(R_i) = \max\left(\frac{bi(R_i)}{\text{total\_}bi}, 1\right) \tag{24}$$

where $bi(R_i)$ is the number of pixels in region R_i which also border on the image, and $total\_bi$ equals the total number of image border pixels.

Thus, for each region in the image, an importance factor rating is calculated for each of the five factors. As already mentioned above, there is little quantitative data which would indicate the relative importance of these different factors, and this relation is likely to change from one image to the next. We have consequently chosen to treat each factor as being of equal importance. However, if it was known that in a particular image fusion applications a particular factor was more important, a weighting of factors could be added easily.

As our aim is to assign a higher importance to areas which rank very strongly in some factors a mean value of all of the importance factors would not be appropriate. Therefore, we have selected to square and sum the factors to produce the final region importance measure (IM)

$$IM(R_i) = \sum_{k=1}^{5} (I_k(R_i))^2 \tag{25}$$

where k sums through the five importance factors. The final IM is produced by scaling the result so that the region of highest importance has an importance value of 1.

D. Measuring the Similarity of Regions Using Visual Information Fidelity

The similarity between the regions in input images and the fused image is determined using the Visual Information Fidelity measure [25]. The natural scene statistics (NSS) model that was used in [25] is the Gaussian scale mixtures (GSM) model in the wavelet domain. Each subband of a scale-space-orientation wavelet decomposition of an image is modeled as a GSM random field (RF). The subband coefficients were partitioned into nonoverlapping blocks of M coefficients each, and block i modeled as the vector C_i. A GSM is a random field (RF) that can be expressed as a product of two different RFs—a GSM $C = C_i \in I$, where I denotes the set of spatial indices for the RF, can be expressed as

$$C = S \cdot U = S_i \cdot \overrightarrow{U}_i : i \in I \tag{26}$$

where $S = S_i : i \in I$ is an RF of positive scalars and $U = U_i : i \in I$ is a Gaussian vector RF with mean zero and covariance $\mathbf{C}_U$. $\overrightarrow{C}_i$ and $\overrightarrow{U}_i$ are M-dimensional vectors and we assume that for the RF U, $\overrightarrow{U}_i$ is independent of $\overrightarrow{U}_j$, $\forall i \neq j$, conditional upon S.

The purpose of a distortion model is to describe how a generic distortion operator disturbs the statistics of an image. The distortion model that was chosen in [25] provides an important functionality while being mathematically tractable and computationally simple. It is a model with a signal attenuation and additive noise

$$D = GC + V = g_i \overrightarrow{C}_i + \overrightarrow{V}_i : i \in I \tag{27}$$

where C denotes the RF from a subband in the reference signal, $D = \overrightarrow{D}_i \in I$ denotes the RF from the corresponding subband from the test (distorted) signal, $G = g_i : i \in I$ is a deterministic scalar gain field, and $V = \overrightarrow{V}_i : i \in I$ is a stationary additive zero-mean Gaussian noise RF with variance $\mathbf{C}_v = \sigma_v^2 \mathbf{I}$. The RF V is white and independent of S and U, whereas the field G is constrained to be slowly varying.

In this algorithm, the HVS is approached as a "distortion channel" that imposes limits on how much information could flow through it. All the sources of HVS are combined into one added noise component that serves as a distortion baseline in comparison to which the distortion added could be evaluated. This combined HVS distortion visual noise is modeled as a stationary, zero-mean, additive white Gaussian noise model in the wavelet domain. Thus, we model the HVS noise in the wavelet domain as stationary RFs $N = \overrightarrow{N}_i : i \in I$ and

$N' = \overrightarrow{N'_i} : i \in I$, where $\overrightarrow{N}_i$ and $\overrightarrow{N'_i}$ are zero-mean uncorrelated multivariate Gaussian with the same dimensionality as $\overrightarrow{C}_i$

$$E = C + N \qquad \text{(reference image)} \tag{28}$$

$$F = D + N' \qquad \text{(test image)} \tag{29}$$

where E and F denote the visual signal at the output of HVS model from the reference and the test images in one wavelet subband, respectively, from which the brain extracts cognitive information. The RFs N and N' are assumed to be independent of U, S and V. We model the covariance of N and N' as

$$\mathbf{C}_N = \mathbf{C}'_N = \sigma_n^2 \mathbf{I} \tag{30}$$

where σ_n^2 is an HVS model parameter (i.e., variance of the visual noise).

If source, distortion and HVS models are described as above, the VIF criterion can be derived [25]. Let $\overrightarrow{C}^N$ denote N elements from C and let S^N, $\overrightarrow{D}^N$, $\overrightarrow{E}^N$ and $\overrightarrow{F}^N$ be defined correspondingly. It is assumed that the model parameters G, σ_v^2 and σ_n^2 are known. The mutual information $I(\overrightarrow{C}^N; \overrightarrow{E}^N)$ measures the amount of information that can be that can be extracted by the human brain when a test image is being viewed. As we would like to measure quality of a particular reference image-test image pair, it is reasonable to "tune" the natural scene model to a specific reference image by calculating $I(\overrightarrow{C}^N; \overrightarrow{E}^N \quad S^N = s^N)$ instead of $I(\overrightarrow{C}^N \overrightarrow{D}^N)$, where s^N denotes a realization of S^N for a particular reference image. All the aforementioned equations stand for one subband analysis, so we are to incorporate multiple subbands by assuming that each subband is completely independent of others in terms of the RFs as well as the distortion model parameters. Thus the final VIF is given by [25]

$$VIF = \frac{\sum_{j \in \text{subbands}} I(\overrightarrow{C}^{N,j}; \overrightarrow{F}^{N,j} | s^{N,j})}{\sum_{j \in \text{subbands}} I(\overrightarrow{C}^{N,j}; \overrightarrow{E}^{N,j} | s^{N,j})} \tag{31}$$

where we sum over the subband of interest and $\overrightarrow{C}^{N,j}$ represents N elements of the RF C_j that describes the coefficients from subband j. VIF is bounded below by zero (when $I(\overrightarrow{C}^N; \overrightarrow{F}^N | s^N) = 0$ and $I(\overrightarrow{C}^N; \overrightarrow{F}^N | s^N) \neq 0$), which indicates that all information about the reference image has been lost in the distortion channel. On the other hand, in case the image is not distorted at all, and VIF is calculated between the reference image and its identical copy, VIF is equal to one.

In our implementation, the VIF given in (31) is computed for a collection of wavelet coefficients from each subband that represents a spatially localized region of subband coefficients for one of the input images and the fused image. Therefore, the total metric value for the input image X is given by

$$M(X,F) = \frac{1}{|R|} \sum_{i=1}^{N_R} VIF\left(R_i^X, R_i^F\right) \cdot IM(R_i^X) \tag{32}$$

and similarly, for input image Y we have

$$M(Y,F) = \frac{1}{|R|} \sum_{i=1}^{N_R} VIF\left(R_i^Y, R_i^F\right) \cdot IM(R_i^Y) \tag{33}$$

where N_R represents the number of regions obtained by the joint segmentation, $|R|$ is the cardinality of the set of all the segmented regions and R_i^X, R_i^Y and R_i^F are the ith region in the image X, Y and F, respectively. The final metric value is equal to the mean value of the $M(X,F)$ and $M(Y,F)$

$$M(X,Y,F) = \frac{1}{2}(M(X,F) + M(Y,F)). \tag{34}$$

IV. Experimental Results

In this section, we demonstrate the benefits of the proposed image fusion metric and perform comparison with the other state-of-the-art metrics [11]–[13] by applying them to evaluate the quality of the image obtained from different fusion schemes. There is a large number of image fusion algorithms available in the literature [5] and four of them are tested here. The focus of this section is to compare different image fusion metrics and not to thoroughly evaluate concepts of the tested image fusion algorithms. The images used in experiments are multimodality surveillance images from TNO Human Factors and medical and multifocus images provided by Rockinger, publicly available at the Image Fusion web site.[1]

We have fused the input images using the following fusion algorithms: the simple averaging method, the ratio pyramid method [26], the fusion method based on the principal component analysis [5] and the DT-CWT image fusion method [27]. In the multiresolution methods (ratio, DT-CWT) input images are decomposed using the given transform and the coefficients of the fused image are computed by choosing the corresponding coefficients of input images with the largest amplitude in the wavelet domain and by averaging the coefficients of lowest resolution. The number of decomposition levels was five. The multiresolution methods usually outperform the simple methods in terms of subjective quality of the fused images, as confirmed by subjective tests in [28], [29]. Ideally, we would want the metric results to be consistent with the subjective quality of the fused image.

For each set of the fused images obtained using these methods we calculate the metric value for the three state-of-the-art metrics and the proposed one. Fig. 2 depicts the two input images from the TNO UN Camp image sequence, the map of regions obtained by the joint segmentation of the two inputs, two maps of the importance of regions in the two modalities and four fused images. It is a good example of two modality inputs which significantly differ and offer complementary information about the surveyed area. There is only one important object in the IR image (the walking person) and a number of important details in the visible modality, such as the fence, the roof of the house and the trees around the house. The details in the visible modality improve situation awareness in the fused image, so ideally more detail in the fused image should be from the visible modality, but

[1][Online.] Available: http://www.imagefusion.org.

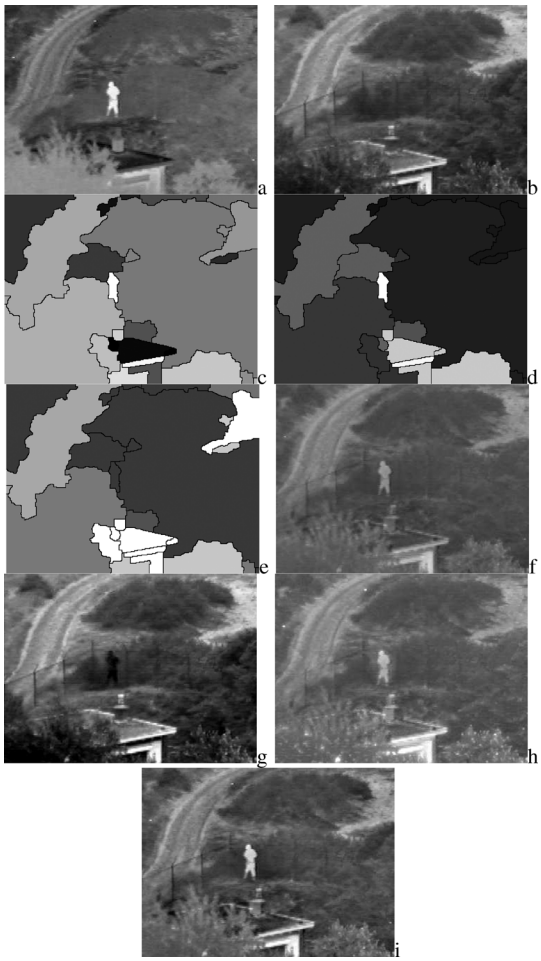

Fig. 2. Visual comparison of fused images generated by the tested image fusion algorithms. (a) input IR image, (b) input visible image, (c) map of regions, joint segmentation, (d) importance of regions in IR input image (brighter color—higher importance), (e) importance of regions in visible input image (brighter color—higher importance), (f) fused image, the simple averaging, (g) fused image, PCA fusion method, (h) fused image, ratio pyramid fusion method, and (i) fused image, DT-CWT fusion method.

TABLE I
PERFORMANCE OF IMAGE FUSION METHODS, MEASURED BY THE IMAGE FUSION METRICS

Metric	Method	UnCamp	Multifocus	Medical
Piella	Average	0.866	0.931	**0.887**
	PCA	0.826	0.930	0.778
	Ratio	0.870	0.902	0.683
	DT-CWT	**0.918**	**0.977**	0.847
Petrovic	Average	0.350	0.586	0.348
	PCA	**0.520**	0.583	0.654
	Ratio	0.412	0.565	0.450
	DT-CWT	0.511	**0.768**	**0.680**
MI	Average	1.065	**1.334**	1.370
	PCA	**1.183**	1.333	**1.441**
	Ratio	1.060	1.304	1.164
	DT-CWT	1.056	1.310	1.098
Proposed	Average	0.563	0.625	0.533
	PCA	0.396	0.683	0.606
	Ratio	0.476	0.607	0.516
	DT-CWT	**0.744**	**0.751**	**0.731**

TABLE II
PERFORMANCE OF IMAGE FUSION METHODS, MEASURED BY THE PROPOSED IMAGE FUSION METRIC AND BY THE PROPOSED METRIC EXCLUDING THE IMPORTANCE MEASURE DATA

Metric	Method	UnCamp	Multifocus	Medical
Proposed	Average	0.563	0.625	0.533
	PCA	0.396	0.683	0.606
	Ratio	0.476	0.607	0.516
	DT-CWT	**0.744**	**0.751**	**0.731**
No IM	Average	0.823	0.814	**0.807**
	PCA	0.834	**0.867**	0.751
	Ratio	0.798	0.854	0.611
	DT-CWT	**0.845**	0.851	0.761

also it is crucial to transfer the walking silhouette to the fused image from the IR input.

The evaluation of the performance of the fusion methods by image fusion metrics is given in Table I. The proposed metric values are compatible to the subjective quality of the fused image – the DT-CWT fusion method clearly produces the fused image with the best contrast and the details well transferred from the input images to the fused one. Piella metric also rates the DT-CWT algorithm as the best performing; however, the dynamic range of the metric values, compared to the proposed metric, is smaller (i.e., the discrimination between different fusion algorithm's performance is inferior). Petrovic metric and the MI metric wrongly rated PCA method as the best performing method, as it clearly fails to include the most important object from the IR image (the walking person) in the fused image.

In addition, we have implemented the proposed algorithm without calculating the importance of regions in the images to be fused, which consequently just calculates the average VIF between the regions in the input images and the corresponding regions in the fused image. Table II gives the comparison of the results obtained when importance map is used in the metric calculation versus the "VIF-only" implementation. It is clear that the modified metric faces the same obstacle as the state-of-the-art metrics, which is to potentially overlook the impact of absence of perceptually important regions in the fused image and therefore its performance is not consistent.

An example of two multifocus input images is given in Fig. 3. One input image has the focus on the clock on the left side and the other input image's focus is on the clock on the right side of the scene. The optimal fused image would include fine detail from both input images, rendering an output with both objects in a clear focus. The evaluation of the four image fusion algorithms' performance is given in Table I. It is clear that all the metrics except the MI metric rate the DT-CWT fusion algorithm as the optimal one, confirming the visual impression of the four fused images.

An example of image fusion in medical imaging is given in Fig. 4, in which a computed tomography (CT) image is fused with a magnetic resonance (MR) image. A well-fused image, with a good contrast and all the important details from both input images incorporated, is required in order to obtain a better diagnostic accuracy. It is clear that the best performance in terms of visual quality of the fused image is obtained when the DT-CWT

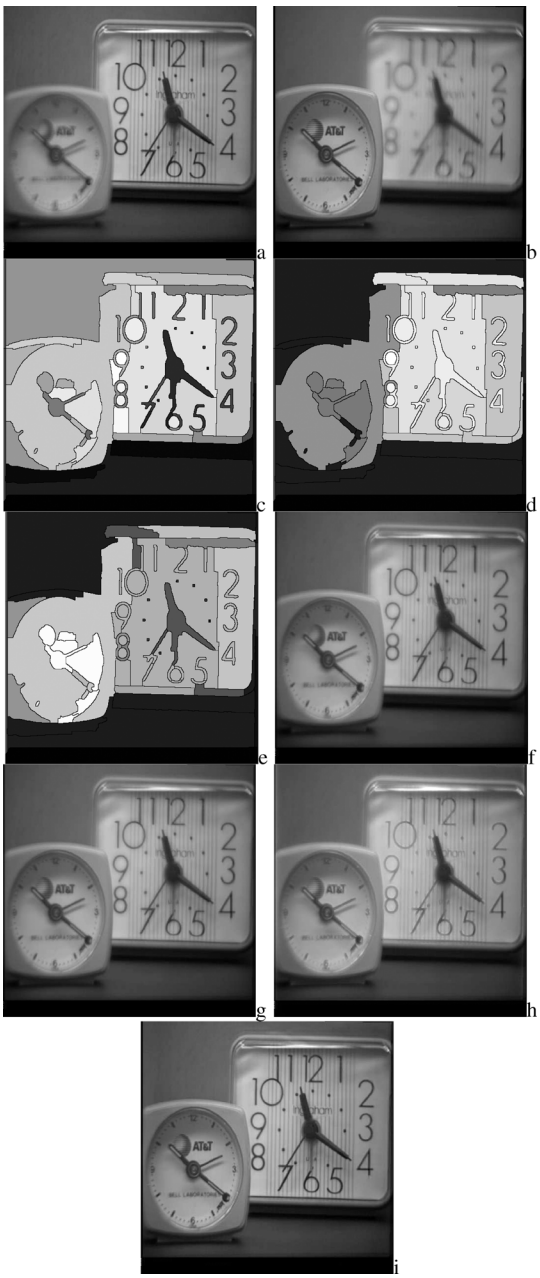

Fig. 3. Visual comparison of fused images generated by the tested image fusion algorithms. (a) input image 1, (b) input image 2, (c) map of regions, joint segmentation, (d) importance of regions in image 1 (brighter color—higher importance), (e) importance of regions in image 2 (brighter color—higher importance), (f) fused image, the simple averaging, (g) fused image, PCA fusion method, (h) fused image, ratio pyramid fusion method, and (i) fused image, DT-CWT fusion method.

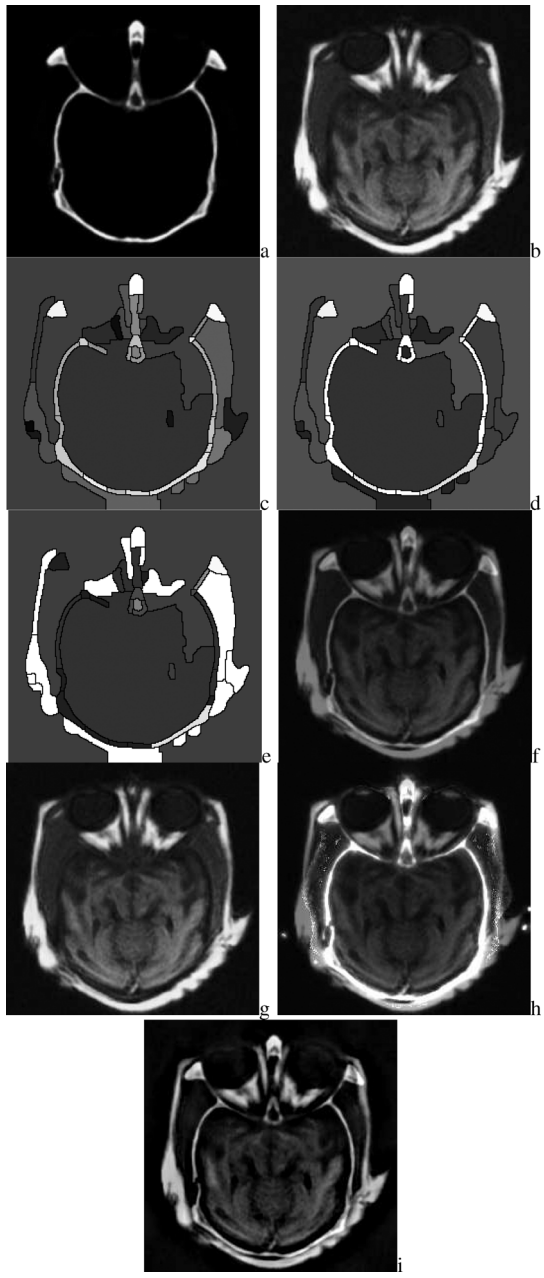

Fig. 4. Visual comparison of fused images generated by the tested image fusion algorithms. (a) input computed tomography (CT) image, (b) input magnetic resonance (MR) image, (c) map of regions, joint segmentation, (d) importance of regions in CT input image (brighter color—higher importance), (e) importance of regions in MR input image (brighter color—higher importance), (f) fused image, the simple averaging, (g) fused image, PCA fusion method, (h) fused image, ratio pyramid fusion method, and (i) fused image, DT-CWT fusion method.

is used. The proposed metric correctly rates the DT-CWT algorithm as clearly the best performing one. This metric result is confirmed with the Petrovic metric, whereas the Piella and MI metrics incorrectly rank the PCA method and simple averaging as the best fusion methods, respectively.

Experimental results show that the proposed metric's ranking of different image fusion methods is consistent with the subjective quality of the fused image. This is confirmed with various examples of images to be fused, ranging from multimodality surveillance image sequences to medical imaging. The existing image fusion metrics (Petrovic, Piella, MI) clearly have less consistent results and the rankings not in accordance with the visual quality of the fused output.

V. Conclusion

This paper presented a novel metric for evaluation of image fusion algorithms, based on the evaluation of similarity of regions in images to be fused with the corresponding regions in the

fused image. The metric uses several factors to quantify the importance of regions in each of the input images, such as contrast, size, and shape of region. The similarity of the corresponding regions in an input image and the fused image is measured using the wavelet-based mutual information measure called visual information fidelity [25].

Tests with a range of images and different fusion algorithms have shown that the proposed metric's rating of fusion algorithms is consistent with the subjective quality of the fused image. The consistency of ranking and dynamic range of the proposed metric outperform the state-of-the-art image fusion metrics, including Piella, Petrovic, and MI metric. This is due to the proposed metric's capability to measure the importance of each of the regions in the input images and accordingly determine how well are the most important features of the input image represented in the fused image. Another important advantage of the proposed metric is that it measures the similarity of higher-level features of an image (such as objects/regions), rather than only measuring the similarity of edges, the mutual information or a window-based similarity.

The proposed image fusion metric has a number of IM and parameters (e.g., $A_{\max}$ equal to 3% of the total image area) which can be tuned to make the metric more suitable to a specific image fusion application. Additionally, the metric is flexible in terms of the number of IMs used. For example, for a specific application in which a trained human operator analyzes the fused image, the IM dependent on the location of a region can be excluded from the calculations, because trained operators usually perform their task by considering the entire fused image.

REFERENCES

[1] H. Maitre and I. Bloch, "Image fusion," *Vistas in Astronomy*, vol. 41, no. 43, pp. 329–335, 1997.

[2] S. Nikolov, P. Hill, D. Bull, and N. Canagarajah, "Wavelets for image fusion," in *Wavelets in Signal and Image Analysis*. Dordrecht, The Netherlands: Kluwer, 2001.

[3] D. Ryan and R. Tinkler, "Night pilotage assessment of image fusion," in *Proc. SPIE*, Orlando, FL, 1995, pp. 50–67.

[4] A. Toet and E. M. Franken, "Perceptual evaluation of different image fusion schemes," *Displays*, vol. 24, no. 1, pp. 25–37, 2003.

[5] G. Piella, "A general framework for multiresolution image fusion: From pixels to regions," *Inf. Fusion*, vol. 4, no. 2003, pp. 259–280, 2003.

[6] H. Li, B. S. Manjunath, and S. K. Mitra, "Multisensor image fusion using the wavelet transform," *Graph. Models Image Process.*, vol. 57, no. 3, pp. 235–245, 1995.

[7] O. Rockinger, "Image sequence fusion using a shift invariant wavelet transform," in *Proc. IEEE Int. Conf. Image Process.*, Washington, DC, 1997, pp. 288–291.

[8] N. Cvejic, D. R. Bull, and C. N. Canagarajah, "Region-based multimodal image fusion using ICA bases," *IEEE Sens. J.*, vol. 7, no. 5, pp. 743–751, May 2007.

[9] V. Petrovic and T. Cootes, "Objectively adaptive image fusion," *Inf. Fusion*, vol. 8, no. 2, pp. 168–176, 2007.

[10] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 2002.

[11] G. Piella and H. Heijmans, "A new quality metric for image fusion," in *Proc. IEEE Int. Conf. Image Process.*, Barcelona, Spain, 2003, pp. 173–176.

[12] V. Petrovic and C. Xydeas, "Objective evaluation of signal-level image fusion performance," *Opt. Eng.*, vol. 44, no. 8, p. 087003, 2005.

[13] G. H. Qu, D. L. Zhang, and P. F. Yan, "Information measure for performance of image fusion," *Electron. Lett.*, vol. 38, no. 7, pp. 313–315, 2002.

[14] N. Cvejic, C. N. Canagarajah, and D. R. Bull, "Image fusion metric based on mutual information and Tsallis entropy," *Electron. Lett.*, vol. 42, no. 11, pp. 626–627, 2006.

[15] E. Niebur and C. Koch, "Computational architectures for attention," in *The Attentive Brain*. Cambridge, MA: MIT, 1997.

[16] A. Yarbus, *Eye Movements and Vision*. New York: Plenum, 1967.

[17] L. Stelmach, W. Tam, and P. Hearty, "Static and dynamic spatial resolution in image coding: An investigation of eye movements," in *Proc. SPIE*, San Jose, CA, 1992, pp. 147–152.

[18] J. Findlay, "The visual stimulus for saccadic eye movement in human observers," *Perception*, vol. 9, pp. 7–21, 1980.

[19] B. L. Cole and P. K. Hughes, "Drivers don't search: They just notice," in *Visual Search*. New York: Taylor & Francis, 1990, pp. 407–417.

[20] A. Gale, "Human response to visual stimuli," in *The Perception of Visual Information*. New York: Springer-Verlag, 1997, pp. 127–147.

[21] G. Elias, G. Sherwin, and J. Wise, "Eye movements while viewing NTSC format television," *SMPTE Psychophys. Subcommittee White Paper*, 1984.

[22] X. Marichal, T. Delmot, V. De Vleeschouwer, and B. Macq, "Automatic detection of interest areas of an image or a sequence of images," in *Proc. IEEE Int. Conf. Image Process.*, Lausanne, Switzerland, 1996, pp. 371–374.

[23] R. J. O'Callaghan and D. R. Bull, "Combined morphological-spectral unsupervised image segmentation," *IEEE Trans. Image Process.*, vol. 14, no. 1, pp. 49–62, Jan. 2005.

[24] W. Osberger and A. J. Maeder, "Automatic identification of perceptually important regions in an image," in *Proc. Int. Conf. Pattern Recognition*, Brisbane, Australia, 1998, pp. 701–704.

[25] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–443, Feb. 2006.

[26] A. Toet, "Image fusion by a ratio of low-pass pyramid," *Pattern Recog. Lett.*, vol. 9, pp. 245–253, 1996.

[27] 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.

[28] T. D. Dixon, J. M. Noyes, T. Troscianko, E. F. Canga, D. R. Bull, and C. N. Canagarajah, "Psychophysical and metric assessment of fused images," in *Proc. Symp. Appl. Percept. Graphics Visualization*, La Coruna, Spain, 2005, pp. 43–50.

[29] T. D. Dixon, E. F. Canga, J. M. Noyes, T. Troscianko, D. R. Bull, and C. N. Canagarajah, "Methods for the assessment of fused images," *ACM Trans. Appl. Percept.*, vol. 3, no. 3, pp. 309–332, 2006.

[30] Y. Chen and R. S. Blum, "A new automated quality assessment algorithm for image fusion," *Image Vision Comput.*, to be published.

[31] H. Chen and P. K. Varshney, "A perceptual quality metric for image fusion based on regional information," *Proc. SPIE*, vol. 5831, no. 34, pp. 34–45, 2005.

[32] P. Soille, *Morphological Image Analysis, Principles and Applications*. Berlin, Germany: Springer-Verlag, 1999.

Nedeljko Cvejic (M'01) received the Dipl.-Ing. degree in electrical engineering from the University of Belgrade, Belgrade, Serbia, in 2000 and the Dr.Tech. degree from the University of Oulu, Oulu, Finland, in 2004.

From 2001 to 2004, he was a Research Scientist at the Department of Electrical and Information Engineering, University of Oulu. From 2005 to 2008, he was a Research Associate with the Department of Electrical and Electronic Engineering of the University of Bristol, Bristol, U.K. Currently, he is a Research Associate with the Signal Processing and Communications Laboratory, Department of Engineering, University of Cambridge, U.K. He has published more than 60 papers and one book. His research interests include image and video fusion, image fusion metrics, sensor networks, and digital watermarking.

Tapio Seppänen received M.Sc. degree in electrical engineering and the Ph.D. degree in computer engineering from the University of Oulu, Oulu, Finland, in 1985 and 1990, respectively.

Currently he is serving as a Professor of Biomedical Engineering at the University of Oulu. He is also the Director of WellTech Oulu Institute at University of Oulu, a member of the board of Infotech Oulu Graduate School, and the Program Director of Biomedical Engineering degree programme, University of Oulu. He teaches and conducts research on multimedia signal processing and biomedical signal processing. Special interests include digital watermarking, pattern recognition applications, and content-based multimedia retrieval. He has contributed to some 300 scientific journal and conference papers.

Simon J. Godsill (M'93) is a Professor of Statistical Signal Processing in the Department of Engineering, University of Cambridge, Cambridge, U.K. He runs a research group in Signal Inference and its Applications, with special interest in Bayesian and statistical methods for signal processing, Monte Carlo algorithms for Bayesian inference, modeling and enhancement of audio and musical signals, tracking, and high-frequency finance. He has published extensively in journals, books, and conferences. He has coedited several journal special issues on topics related to Monte Carlo methods and Bayesian inference.

Prof. Godsill was an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING and on the IEEE Signal Processing Theory and Methods Committee. He has been on the scientific committees of numerous international conferences, and workshops and he has organized special sessions ranging from audio processing topics to probability and inference.