1 Introduction

Image fusion is the technique of combining information from multiple images of the same scene, but either captured from different sensors, different viewpoints or different times. The fused result is a single image containing all important features of each original image thereby being more comprehensive and more suitable for further human operators or further machine processing tasks. A number of applications employ image fusion since they need complementary spatial and spectral information in a single image, but a capability of a single sensor used in a single moment is limited by design or observational constraints. Multiple images are therefore acquired in different manners. Several source modalities are utilised to exhibit various characteristics, such as type of degradation, texture properties, colours, spectral bands, etc. These applications include satellite imaging, medical imaging, robot vision, monitoring and surveillance. Multiview images are taken from the same modality, same time but from different positions to create a 3D image. Multifocus images are taken repeatedly with various focal lengths to create one sharp image. Same camera at the same position captures images at different times is used to detect changes, to enhance nighttime scenery, or to restore an image from blurred and noisy image sequence. Multi-exposure is the superimposition of images captured from various exposures to create a single image used for context enhancement or for increasing dynamic range.

The fusion process can take place at different levels of information representation [1], i.e. signal, pixel, feature and symbol level. **Signal-level** fusion combines several signals that have the same general format. At **pixel-level** (or image-level in some literature), images are combined by considering individual pixel values or small arbitrary regions of pixels in order to make the fusion decision, while at **feature-level**, the fusion process is decided based on characteristics of a set of regions in each input image, such as shape, contrast and texture. This has advantages over pixel-based methods as more intelligent semantic fusion rules can be considered based on actual features in the image, rather than on single or arbitrary groups of pixels. Fusion at **symbol-level** allows the information to be effectively combined at the highest level of abstraction. The input images are usually processed individually for information extraction and classification. This results in a number of symbolic representations which are then fused according to decision rules. The example system involving different types of multisensor fusion is demonstrated in Fig. 1. It is obvious that the higher the fusion level, the more the complexity.

To enhance interpretation for human observer, some generic requirements should be imposed on the fusion result [3]: (i) the fused image should preserve as closely as possible all relevant information contained in the input images; (ii) the fusion process should not introduce any artefacts or inconsistencies, which can distract or mislead the human observer, or any subsequent image
This review report comprises an extensive literature survey of state-of-the-art for image fusion. To achieve the best performance of image fusion, particularly in the pixel level, an image registration is required to enable the input images to spatial alignment. This will be briefly reviewed in Section 2. The pixel-level fusion is discussed in Section 3, followed by the feature-level fusion in Section 4. Finally, conclusions and some recommendations are present in Section 6.

2 Image Registration

Image registration is the process of establishing point-by-point correspondence between a number of images, describing the same scene. There are two types of image registration used in image fusion manner, single- and multi-modality methods. Single-modality methods register images in the same modality acquired by the same sensor type, while multi-modality registration methods register images acquired by different sensor types. Depending on the types of images, the registration can be produced with either global transformation, i.e. rotation, translation, scaling, affine and projective, or ‘nonrigid’ transformation of which it is capable of local warping.

2.1 Single-Modality Method

Early work employed full-frame alignment, where camera translation and rotation are modelled by orthographic projection [4]. This method does not give impressive results as the image always combines a variety of depths which have different translations. A local block-based flow has therefore been proposed [5]. These approaches employ the intensity of the images. The best match occurs when the registration process achieves the minimum intensity difference between the reference image and the input images. The other methods find correspondence between image features such as points, lines, and contours [6]. For nonrigid transformation, theoretically, the warping
process is finished when the iterative process reaches the minimum of the cost function, which can be computed from intensity-based or feature-based technique. The review of the geometric transformations for nonrigid body registration can be found in [7]. A hierarchical image registration technique decomposes the nonrigid matching problem into an elastic interpolation of numerous local rigid registrations of sub-images [8]. Motion estimation is performed iteratively, firstly by using coarser level coefficients to determine large motion components and then by employing finer level coefficients to refine the motion field. Recently, the dual-tree complex wavelet transform (DT-CWT) is utilised for hierarchical image registration [9]. This algorithm is based on phase-based multidimensional volume registration, which is robust to noise and temporal intensity variations.

2.2 Multi-Modality Method

Various applications of image fusion use multiple source modalities in which a sum of squared intensity differences cannot possibly be applied. Therefore, other image similarity measures are used. One of the most popular measures is mutual information (MI) [10]. It computes from the joint intensity histogram of two different images. The technique attempts to find the registration by maximising the information that one image provides about the other. The main drawback of the MI is that it is highly non-convex and has typically many local maxima, so the normalised intensity gradients are therefore employed in [11]. To improve the performance for registering the noisy image showing no clear image structure, the cross correlation is introduced to used with the MI in the hierarchical registration [12]. Some registration approaches for medical imaging use learning similarity functions as the exact registration is known and it can then be used to improve the existing algorithms [13].

3 Pixel-Level Fusion

Pixel-level fusion methods are computationally efficient and easy to implement as the original measured quantities are directly involved in the fusion process. A number of pixel-level fusion methods have therefore been developed. They can be roughly divided into two groups, spatial image fusion methods and transform image fusion methods.

3.1 Spatial Image Fusion

The simplest approach of the image fusion employs the weighted average of the pixel intensity of the corresponding pixels of source images. The registered input images $I_1, I_2, \ldots, I_N$ are fused using the fusion rule $\phi$ to create a new image $I_F$ as follows.

$$I_F = \phi(I_1, I_2, \ldots, I_N) = \alpha_1 I_1 + \alpha_2 I_2 + \ldots + \alpha_N I_N$$

where $\sum_{n=1}^{N} \alpha_n = 1$. The weightings can be obtained from the eigenvectors which keep the key features in the original image and reduces noise level, called the principal component analysis (PCA) method [14,15], or can be computed adaptively using local information, such as variance and colour temperature. A major problem when fusing images by averaging is that the contrast of the image features uniquely presented in only one of the images is reduced.
Another spatial image fusion method, which is sometimes used in medical image interpretation [16, 17], is the checker-board display. This method is based on the ability of the human eye to quickly and easily integrate visual information. One image is displayed in the ‘white’ squares and the other is displayed in the ‘black’ squares. The major drawback is that the resolution of each of the input images is effectively reduced by a factor of two.

Recently, the gradient field was proposed in an image fusion for context enhancement [18, 19]. The input images are combined using weighted combination of gradient magnitude. The results show that operating in the gradient domain can better preserve local detail while avoiding visible artefacts. This method is generally used for merging multi-exposure images. Examples of multi-exposure fusion are shown in Fig. 2 (d)-(g) which were generated using three multi-exposure images (a)-(c) [19].

### 3.2 Transform-domain Image Fusion

Transformations can expose some invisible useful information. Some image fusion techniques therefore initially transform the original images into another domain by applying a forward transformation $\omega$ and then combine transformed image. The inverse transformation $\omega^{-1}$ is applied to create the final result as demonstrated in Eq. 2.

$$I_F = \omega^{-1}(\phi(\omega(I_1), \omega(I_2), \ldots, \omega(I_N)))$$

---

Figure 2: (a)-(c) The multi-exposure images. Results of fusion methods proposed in (d) [20], (e) [21], (f) [22] and (g) [19]
Transform based approaches provide a fused image with full contrast, but these approaches are sensitive to sensor noise. Two transforms most often used in image fusion are image pyramids and wavelet transforms. Other multi-resolution transforms, such as curvelet (CVT) [23,24] and contourlet (CT) [25,26], have been reported worse performances [27]. The nonsubsampled version of contourlet (NSCT) improves fusion results significantly [28,29], but it is the combination of Laplacian pyramid and overcomplete DWT.

Pyramid decomposition is the earliest multiscale transform used for image fusion. The general idea of all pyramid based fusion schemes is that a multi-resolution image pyramid is constructed from each of the input images. Then, a composite image pyramid is formed from the input image pyramids by applying some fusion rule. Once the fused image pyramid is calculated, the fused image is constructed from it. Two pyramid decomposition schemes are used in general, i.e. Laplacian and Gaussian pyramids, but they make use of different fusion rules, such as contrast pyramid [30], gradient pyramid [31], morphological pyramid [32], etc.

It can be noted here that the experimental results indicate that the appropriate setting for the number of decomposition levels is ‘four’ [27,33]. However, it also depends on the resolution of the original images. As the experiment in [34], the larger the resolution ratio is the more decomposition levels are required to achieve a satisfactory fusion result. It is a trade-off between the capability of catching spatial details and the sensitivity to noise and transform artefacts. When the number of decomposition levels is too large, one coefficient in coarse resolutions responds to a large group of pixels of fused image. Therefore, an error in coarse resolutions has a great effect on final fused image. Some errors inevitably occur in the process of fusion, producing some artificial distortion. When the number of decomposition levels is too small, spatial details cannot be captured very well.

### 3.2.1 Wavelet Transform

Wavelet transforms have been successfully used in many fusion schemes. A common wavelet decomposition used for fusion is the discrete wavelet transform (DWT). It has been found to have some advantages over pyramid schemes [35,36], such as increased directional information, less blocking artefacts, better signal-to-noise ratios, improved perception when compared using human analysis. A major problem of the DWT is its shift variant nature caused by sub-sampling which occurs at each level. A small shift in the input signal results in a completely different distribution of energy between DWT coefficients at difference scales [37]. A shift invariant DWT (SIDWT) yields a very over-complete signal representation as there is no sub-sampling thereby causing high complexity [38].

Dual-Tree Complex Wavelet Transform (DT-CWT) is an over-complete wavelet transform that provides near shift-invariance and better directional selectivity, with less increased memory and computational cost compared to the SIDWT [39]. It employs two fully decimated trees, one for the odd samples and one for the even samples generated at the first level. This increases directional sensitivity over the DWT and is able to distinguish between positive and negative orientations giving six distinct sub-bands at each level, the orientations of which are $\pm 15^\circ$, $\pm 45^\circ$, $\pm 75^\circ$. These properties of the DT-CWT give improved fusion results over the traditional DWT and most other fusion schemes [3]. The example of wavelet-based image fusion using different transformations is shown in Fig. 4 [40]. The result of the DWT shows obvious ringing artefact and the result of the DT-CWT is sharpest. In [41], the authors use task-based fused image quality to indicate that the DT-CWT and SIDWT led the better performance than the DWT and contrast pyramid for
Another advantage of the DT-CWT is that the phase of the complex wavelet coefficients corresponds to the precise location of directional features in its support regions. Therefore slightly misaligned images after image registration process can simply be corrected by adjusting the the DT-CWT coefficients, $d^{\theta,l}(x,y)$, of the fused image with a unit vector representing the average phase from all frames, $N$, as shown in Eq. 3. This is employed to remove rippling effect in the heat haze problem [42].

$$
\tilde{d}^{\theta,l}(x,y) = \frac{\sum_{n=1}^{N} d^{\theta,l}_n(x,y)}{\sum_{n=1}^{N} |d^{\theta,l}_n(x,y)|} \left|d^{\theta,l}(x,y)\right|
$$

The pixel-level fusion schemes employ the wavelet transform to obtain a multiresolution (MR) decomposition of the input images. The wavelet coefficients are combined using some fusion rules to produce a single set of coefficients corresponding to the fused image.

### 3.2.2 Generalized Intensity-Hue-Saturation transformation

Generally when fusing visual colour image with different modalities, the colour image (RGB) is transformed to intensity and colour components, e.g. HSV, YCbCr, and IHS. The intensity (V, Y or I) is then used in fusion process and the colour components are later combined for converting back to RGB space for final results. However, the simple fusion process may not work well for merging multispectral (MS) imagery of lower resolution with a higher resolution image, such as in the remote sensing applications. This is because the colour transform can possibly suffer from individual colour
distortion on saturation compression. Example of this effect is show in Fig. 5 [43]. It also clearly shows that the wavelet approach (Fig. 5 (d) and (e)) preserved the spectral characteristics of the multispectral image better than the IHS method. In general, however, images fused by wavelets have much less spatial information than those fused by the IHS method and it is not efficient enough to merge massive volumes of data from new satellite images. The Generalized Intensity-Hue-Saturation (GIHS) transformation, first introduced in [44], has been used for these applications [45].

A pan-sharpening algorithm is employed with a genetic algorithm which is adopted to define the injection model to establish how the missing high-pass information is extracted from the original high-resolution image (e.g. panchromatic images).

### 3.2.3 Independent Component Analysis

Independent component analysis (ICA) bases are first employed to perform image fusion between noisy and noiseless dataset [46]. The ICA bases are trained using images of similar context. The input images for fusion are firstly devided into image patches, which are subsequently transformed using the ICA. The transformed images of the noisy images are then denoised via sparse code shrinkage [47]. The denoised results are combined information in the ICA bases context and, finally, they are inversed transform to a fused image. Note that an adaptive fusion rule is used in [46]. The results show that the ICA based method outperforms the wavelet-based method, however, with slightly increased computational complexity. The combination of the wavelet transform and the ICA has been proposed for medical image fusion in [48]. An example of multi-modal image fusion using wavelet, dual-tree wavelet and ICA is shown in Fig. 6 (d)-(f), respectively [46].

### 3.2.4 Fusion Rules

A number of fusion rules have been proposed for pixel-based fusion methods. The common rule is *averaging* which is generally used only at the low-pass coefficients since the average of the high-pass...
sub-bands tends to blur images and reduce the contrast of features appearing only in one image. Some of fusion rules applied to high-pass sub-bands are summarised below:

- **Maximum selection**: The simplest and efficient fusion rule was initially suggested in [3]. The maximum-selection scheme selects the largest absolute wavelet coefficient at each location from the input images as the coefficient at that location in the fused image, as wavelets tend to pick out the salient features of an image. The maximum-selection results fulfill most of the requirements mentioned in Section 1. Relevant information is generally preserved in the fused image and few noticeable artefacts are introduced.

- **Weighted average**: The resultant coefficient for reconstruction is calculated from some measures via a weighted average of input images coefficients. For example, in [31], the normalised correlation between the two images sub-bands over a small local area is calculated. The results are slightly improved compared to those of the maximum-selection rule in term of noise reduction, but the sharpness may be diminished.

- **Adaptive weighted average**: Weighted average is applied locally to merge multiple coefficients. A decision map is created based on the activity of an arbitrary block around a central coefficient [49]. An area based selection rule with consistency verification is employed in [35]. If the activity of pixels from different images is similar, an average could be considered. Finally consistency check is made. In [46], the $L_1$-norm is used to detect active regions and constant background where the weighted combination and the mean rule are applied, respectively. A combination of measures can be used to create a weight map. In [20], the contrast, saturation and well-exposedness are employed in Laplacian and Gaussian Pyramid.
based transformations for multiple exposure fusion process.

- **Model-based weighted average**: As wavelet coefficients exhibit non-Gaussian characteristics, the families of generalised Gaussian distribution (GGD) and symmetric alpha-stable distribution (SαS) are used for modelling image wavelet coefficients [50–52]. The fusion results of visible/IR images show improvement over convolutional weighted average. Cauchy distribution is used for context enhancement fusion in [53]. Meridian distribution is employed in [54] for medical images. The weights of two later techniques are optimised via Maximum Likelihood (ML) estimation.

- **Denoising**: perform simultaneous fusion and denoising by thresholding the transform’s coefficients [55] or using the local average modulus of gradients [56]. The shrinkage function is combined with the model-based weighted average using bivariate Laplacian-based and bivariate Cauchy-based techniques [52]. This method produces great fused results in many applications, including medical imaging and other multiband sensors.

- **Replacing**: This method is used when the different resolution images are fused. Generally the lower-resolution images contain more dimensions, such as colours and spectrum, while the high-resolution image is sharper. Therefore the images are transformed into common component (intensity or the first principal component image) and such component of the high-resolution image is substituted to that of the lower-resolution images [57].

### 4 Feature-Level Fusion

The majority of application of a fusion scheme are interested in features within the image, not in the actual pixels. Therefore, it seems reasonable to incorporate feature information into the fusion process [49]. There are a number of perceived advantages of this, including [58]:

- **Intelligent fusion rules**: Fusion rules are based on combining groups of pixels which form the regions of an image. Thus, more useful tests for choosing the regions of a fused image, based on various properties of a region, can be implemented.

- **Highlight features**: Regions with certain properties can be either accentuated or attenuated in the fused image depending on a variety of the regions’ characteristics.

- **Reduced sensitivity to noise**: Processing semantic regions rather than at individual pixels or arbitrary regions can help overcome some of the problems with pixel-fusion methods such as noise, blur and mis-registration.

- **Registration and video fusion**: The feature information extracted from the input images, could be used to aid the registration of these images. Region-video fusion schemes could use motion estimation to track the fused features, allowing the majority of frames to be quickly predicted from some fully fused frames.

A method of achieving feature-level fusion is with a region-based fusion scheme. The process initialises using image segmentation to produce a set of homogeneous and meaningful regions. Ideally, all features of the image are represented by single separate regions. Various properties of these regions can be calculated and used to determine which features from which images are included in the fused image.
4.1 Segmentation

Segmentation algorithms can be used to extract regions of similar properties each ideally corresponding to a feature or features in the image. The algorithms partition the image into some non-intersecting regions so that each region is homogeneous and the union of no two adjacent regions is homogeneous [59]. The results of segmentation directly affect the fusion process; therefore, the segmentation algorithm should have the following properties [58].

- **Features segmented as single separate regions**: if a feature is missed, it may not be included in the fused image. If a feature is split into more than one region, each will be treated separately, possibly introducing artefacts into the fused image.

- **As few regions as possible**: the time taken to compute the fused image increases with the number of regions.

A number of segmentation methods have been proposed over decades. They can be broadly divided into several groups:

- **Edge-Based**: Locate boundaries at discontinuities of the image. While this emulates human perception, often these methods will fail to close these boundaries. Secondly, image intensity is not a good basis for segmenting a textured image.

- **Region-growing**: Always generate closed connected regions. An example of this is watershed transformation, which uses gradient information in a region-growing method. Due to noise in the gradient image, this method tends to over-segmentation the images.

- **Clustering**: Rather than the local nature of the methods above, clustering takes a more global view by optimising a cost function over the whole image. Pixels or small regions are grouped together using knowledge not available at a local level and can result in a perceptually better segmentation.

4.1.1 The CMSUIS Segmentation Algorithm

An adapted version of the combined morphological-spectral unsupervised image segmentation algorithm (CMSUIS) has been developed to enable it to handle sets of multi-modal images. This algorithm was developed at the University of Bristol and is described in [59].

The algorithm works in two main stages. Firstly, both texture and intensity information are used to produce an initial segmentation, consisting of regions of similar intensity or texture. Textual information is extracted from the sub-bands of a DT-CWT. A perceptual gradient function, is derived from this textual and intensity information in the image. A watershed transform is used to provide as initial segmentation from the gradient function. This initial segmentation is an over-segmentation of the image. A watershed algorithm defines closed connected regions with an absence of edge energy (a lack of gradient in this instance). The second stage groups together these primitive regions using a spectral clustering technique. In other words, regions thought to be representing the same feature (because of similarities between the regions) are joined together.
The algorithm employs both texture and intensity information which should be the best for segmenting natural images in the visible spectrum. This might not be true for the different modalities, such as IR images which tend to lack of textual information, but contain strong intensity difference between regions. Therefore a join segmentation was introduced [60,61]. When fusing different modalities, a weak region in one image may correspond to a strong region in another image. There is a potential advantage of using information from all images of a scene to produce single segmentation map for all images in the set.

In general, jointly segmented images work better for fusion. This is because the segmentation map will contain a minimum number of regions 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 This may cause artefacts at the fused result. However, if the information from the segmentation process is going to be used to register the images or if input images are completely different, it can be useful to separately segment the images as the overlapped region may contain different features.

4.2 Region-based Fusion with DT-CWT

Due to its shift invariance, orientation selectivity and multiscale properties, the DT-CWT is widely used in image fusion where useful information from a number of source images are selected and combined into a new image [61–63]. Region-based fusion methods have been introduced by firstly segmenting N images individually or jointly [49,61,62]. The segmentation map $S_n$ of each image is down sampled by 2 to give a decimated segmentation map $S_{n\theta,l}, n \in N$ of level $l$ and sub-band $\theta$ of the DT-CWT representation, where $\theta \in (1,\ldots,6)$. If a list of all regions, $T_n$, of image $n$ is $R_n = \{r_{n,1}, r_{n,2}, \ldots, r_{n,T_n}\}$, a multi-resolution priority map $P_n$ is generated as

$$P_n = \{p_{n,r_{n,1}}, p_{n,r_{n,2}}, \ldots, p_{n,r_{n,T_n}}\}$$

for each region in each image $n$. Regions are then either chosen or discarded based on this priority and the fusion rule, $\phi$, to give the wavelet coefficients of the fused image. A mask, $M$, is generated, where:

$$M_t = \phi (p_{1,t}, p_{2,t}, \ldots, p_{N,t})$$

The mask is the same size as that of the wavelet coefficient region in the fused image. The algorithm always chooses the region with the maximum priority to determine which image each of the coefficients representing a region, $t$, should come from. If $S_i \neq S_j$, a segmentation map, $S_F$, is created such that $S_F = S_1 \cup S_2 \cup \ldots S_N$. Thus, where two regions $r_{i,p}$ and $r_{j,q}$ from image $i$ and $j$ overlap, both will be split into two regions, each with the same priority as the original. Finally, the fusion image is obtained by performing the inverse transform on the fused wavelet coefficients.

4.2.1 Fusion Rules

The lowpass wavelet coefficients of the fused image are simply constructed from the average of the lowpass values of all images, while the highpass coefficients are selected according to an activity
measurement which motivated by the fact that the human visual system is primarily sensitive to local contrast changes, e.g. edges or corners.

Since wavelet coefficients having large absolute values contain the information about the salient features of the images such as edges and lines, a good fusion rule is to take the maximum of the absolute values of the corresponding wavelet coefficients of the wavelet coefficients over a region. Alternatively, the variance or the entropy of the wavelet coefficients can be used. The priority \( P \) of region \( r_{n}^{\theta} \in R \) in image \( n \) is computed with the detail coefficients \( d_{n}^{\theta,l}(x,y) \) of level \( l \) and sub-band \( \theta \), where \( |r_{n}^{\theta}| \) is the size of such area used for normalisation. The mask \( M_{r_{n}^{\theta}} \) is then generated from ranked priority to construct the fused image.

- **Maximum absolute value**: \( P(r_{n}) = \frac{1}{|r_{n}|} \sum_{\forall \theta, \forall l, (x,y) \in r_{n}} |d_{n}(\theta,l)(x,y)| \)

- **Variance**: \( P(r_{n}) = \frac{1}{|r_{n}|} \sum_{\forall \theta, \forall l, (x,y) \in r_{n}} (d_{n}(\theta,l)(x,y) - \bar{d}_{n}(\theta,l))^{2} \), where \( \bar{d}_{n}(\theta,l) \) is the average wavelet coefficient of the region \( r_{n} \)

- **Normalised Shannon Entropy**: \( P(r_{n}) = \frac{1}{|r_{n}|} \sum_{\forall \theta, \forall l, (x,y) \in r_{n}} d_{n}^{2}(\theta,l)(x,y) \log d_{n}^{2}(\theta,l)(x,y) \)

The model-based weighted average has also been used in the region-based image fusion. In [63], the region features are modelled using bivariate alpha-stable distributions (BoS) and the statistical measure of similarity between corresponding regions of the source images is calculated as the Kullback-Leibler distance (KLD) between the estimated BoS model. The results of the visible/IR image fusion and MRI/CT image fusion show that this algorithm achieves better performance than other fusion rules. Example of region-based image fusion is shown in Fig. 7.

5 Fused Image Assessment

The assessment of fused images has broadly been carried out in three ways: through the use of subjective ratings of quality; through the use of computational metrics that model image quality on
the basis of some aspects of the human visual system; and using objective human tasks to obtain data about the usefulness or reliability of the given fusion system. In most scenario, the use of subjective ratings tasks is not viable, as they are particularly time-consuming and troublesome with video assessment processes. Moreover, it is worth noting that subjective ratings have been shown to lead to differing patterns of results from objective human tasks, making their value in real terms questionable [66]. Therefore, in this review, the objective and the task-based assessments are stated below.

5.1 Fused Image Metrics

Several approached to fused image quality evaluation exist. These include qualitative tests with human participants and quantitative or objective tests. A number of image quality metrics have been proposed including mean square error (MSE), root mean square error (RMSE), peak signal to noise ration (PSNR), mean absolute error (MAE) and quality index. All of these require a reference image, which is usually the ideal fused image. However, in practice, such an ideal fused image is rarely known. Hence, other fused image metrics such as mutual information (MI) [67], the Petrovic and Xydeas metric [68] and Piella’s Quality Index [69] have been proposed. These estimate how and what information is transferred from the input images to the fused image.

5.1.1 Mutual Information

Mutual information has emerged as an alternative to RMSE [67]. The MI measures the degree of dependence of the two random variable A and B. It is defined by Kullback-Leibler measure.

\[ I_{AB}(a,b) = \sum_{a,b} p_{AB}(a,b) \cdot \log \frac{p_{AB}(a,b)}{p_A(a) \cdot p_B(b)} \] (6)

where \( p_{AB}(a,b) \) is the joint distribution and \( p_A(a) \cdot p_B(b) \) is the distribution in the case of complete independence. Considering two input images A, B and a new fused image F, the image fusion performance measure can be defined as:

\[ M_{AB}^F = \frac{I_{FA}(f,a) + I_{FB}(f,b)}{2} \] (7)

5.1.2 Petrovic and Xydeas Metric

Petrovic and Xydeas Metric measures the amount of edge information ‘transferred’ from the source image to the fused image to give an estimation of the performance of the fusion algorithm [68]. It used a Sobel edge operator to calculate the strength \( g(n,m) \) and orientation \( \alpha(n,m) \) information of each pixel in the input and output images. The relative strength and orientation ‘change’ values, \( G_{AF}(n,m) \) and \( A_{AF}(n,m) \) of an input image A with respect to the fused image F, are defined as:

\[ G_{AF}(n,m) = \begin{cases} \frac{g_F(n,m)}{g_A(n,m)}, & \text{if } g_A(n,m) > g_F(n,m) \\ \frac{g_A(n,m)}{g_F(n,m)}, & \text{otherwise} \end{cases} \] (8)
\[ A^{AF}(n,m) = \frac{|\alpha_A(n,m) - \alpha_F(n,m)| - \pi/2}{\pi/2} \]  

These measures are then used to estimate the edge strength and orientation preservation values, \( Q^A_{gF}(n,m) \) and \( Q^A_{\alpha}(n,m) \):

\[ Q^A_{gF}(n,m) = \frac{\Gamma_g}{1 + e^{k_g(G^{AF}(n,m) - \sigma_g)}} \]  
\[ Q^A_{\alpha}(n,m) = \frac{\Gamma_\alpha}{1 + e^{k_\alpha(G^{AF}(n,m) - \sigma_\alpha)}} \]

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

\[ Q^{AF}(n,m) = Q^{A_{gF}}(n,m) \cdot Q^{A_{\alpha}}(n,m), \quad 0 \leq Q^{AF}(n,m) \leq 1 \]

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

\[ Q^{AB/F}_{p}(n,m) = \frac{\sum_{n=1}^{N} \sum_{m=1}^{M} Q^{AF}(n,m)w_A(n,m) + Q^{BF}(n,m)w_B(n,m)}{\sum_{n=1}^{N} \sum_{m=1}^{M} w_A(n,m) + w_B(n,m)} \]

It can be observed that the edge preservation value \( Q^{AF}(n,m) \) 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 whilst the region information is ignored.

5.1.3 Piella Metric

This image fusion quality index (IFQI) measures three different aspects: correlation, luminance distortion and contrast distortion, which was introduced by Wang and Bovik [70]. It is defined as:

\[ Q = \frac{4\sigma_{xy}}{(\sigma_x^2 + \sigma_y^2)(\bar{x}^2 + \bar{y}^2)} \]

where \( \bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \), \( \bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i \), \( \sigma_x^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2 \), \( \sigma_y^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{y})^2 \) and \( \sigma_{xy} = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y}) \). To understand the meaning of \( Q \), it can be decomposed as a product of three components:

\[ Q = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \cdot \frac{2 \bar{x} \bar{y}}{\bar{x}^2 + \bar{y}^2} \cdot \frac{2 \sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2} \]
The first component is the correlation coefficient between \( x \) and \( y \). The second component corresponds to the luminance distortion and the third factor measures the contrast distortion. The maximum value of \( Q = 1 \) is achieved when \( x \) and \( y \) are identical.

In order to apply this metric for image fusion evaluation, Piella and Heijmans [69] introduce salient information to reflect the relative importance of image \( A \) compared to image \( B \) within the window \( W \).

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

(16)

where \( F \) is the fused image and \( c(w) \) is the overall salience of a window. \( \lambda(w) = \frac{s(A|w)}{s(A|w) + s(B|w)} \), where \( s(A|w) \) and \( s(B|w) \) are the local saliencies of image \( A \) and \( B \), respectively.

Finally, to take into account some aspect of the human visual system (HVS) which is the relevance of edge information, the same measure is computed with the ‘edge images’ \( (A', B', \text{and} \ F') \) instead of the greyscale images \( A, B \) and \( F \).

\[
Q_E(A, B, F) = Q_w(A, B, F)^{1-\alpha} \cdot Q_w(A', B', F')^{\alpha}
\]  

(17)

where \( \alpha \) is a parameter that expresses the contribution of the edge images compared to the original image. As with the previous metrics, this metric does not require a groundtruth or reference image.

5.2 Task-Based Fused Image Assessment

It is essential to consider what task is being undertaken when attempting to assess the quality of a fused image. Image fusion can be used to achieve various ends, such as to enhance spatial, spectral or temporal resolution. It is therefore imperative that the task of the fusion process is defined in advance of the quality metric used. As the output of a fusion process might be fed directly back into the system without reference to a human viewer, this could have important applications on what metric is relevant. Likewise, the kind of processing used on the original data can alter spectral content, or cause blurring if images with low Signal-to-Noise Ratio (SNR) are fused, and must therefore be taken into account [71].

The task-based assessment applies a task to each image-rating scenario, with the advantage being that objective data is drawn from the sample, in the form of accuracy and reaction times [41]. This data is of more meaningful (less arbitrary) value and actual psychological processes can be more easily inferred from it. In addition, if the correct task is chosen, then the results have direct application to a given problem. This assessment method use the psychophysical testing combined with subjective quality and metric assessment. More complex method employs a scanparth analysis with the use of an eye-tracking paradigm when assessing video content [66].

6 Conclusions

The review of state-of-the-art algorithms for image fusion at pixel and feature levels is presented here. Both fusion levels can be processed either in spatial domain or in transform domain. In
many applications, particularly those of multi-modalities, image fusion algorithms implemented un
taxform domain show better results as they enable the use of salient information that human visual system is sensitive to. Among these transform methods, the DT-CWT seems to have the best performance because of its properties of good shift invariance and directional selectivity, while its complexity is not high. Numerous fusion rules have been introduced for general purpose or for specific applications. The maximum selection is the simplest and effective decision for most applications, but it is extremely sensitive to noise. An adaptive weighted average using statistical modelling, e.g. non-Gaussian, is possibly the best fusion rule for the DT-CWT. It can be applied to various applications, such as visible/IR imaging, context enhancement, and medical imaging, both with noisy and noiseless images.

This review also presents two methods of fused image assessent. Objective quality metrics, including Mutual Information, Petrovic and Xydeas Metric and Piella Metric, are utilised to assess the fused image without the knowledge of groundtruth. As each image fusion technique has been used in a variety of applications, their respective results should be assessed depending on the tasks for which they are used. The task-based fused image assessment is therefore mentioned in this review.

References


