Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®
Information Fusion 4 (2003) 259–280
INFORMATION FUSION
www.elsevier.com/locate/inffus

A general framework for multiresolution image fusion: from pixels to regions ☆

Gemma Piella *

Center for Mathematics and Computer Science (CWI), P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

Received 14 May 2002; received in revised form 17 March 2003; accepted 1 April 2003

Abstract

This paper presents an overview on image fusion techniques using multiresolution decompositions. The aim is twofold: (i) to reframe the multiresolution-based fusion methodology into a common formalism and, within this framework, (ii) to develop a new region-based approach which combines aspects of both object and pixel-level fusion. To this end, we first present a general framework which encompasses most of the existing multiresolution-based fusion schemes and provides freedom to create new ones. Then, we extend this framework to allow a region-based fusion approach. The basic idea is to make a multiresolution segmentation based on all different input images and to use this segmentation to guide the fusion process. Performance assessment is also addressed and future directions and open problems are discussed as well.

Keywords: Image fusion; Multiresolution decompositions; Multisource segmentation; Region-based fusion

1. Introduction

Extraordinary advances in sensor technology, microelectronics and communications have brought a need for processing techniques that can effectively combine information from different sources into a single composite for interpretation. In image-based application fields, image fusion [1] has emerged as a promising research area.

In this paper we are concerned with the fusion of visual information. Indeed, as many sources produce images, image processing has become one of the most important domains for fusion. Image fusion can be broadly defined as the process of combining multiple input images into a smaller collection of images, usually a single one, which contains the 'relevant' information from the inputs, in order to enable a good understanding of the scene, not only in terms of position and geometry, but more importantly, in terms of semantic interpretation. In this context, the word 'relevant' should be considered in the sense of 'relevant with respect to the task the fused images will be subject to', in most cases high-level tasks such as interpretation or classification. In the sequel, we will refer to this 'relevant' information as *salient* information. The images to be combined will be referred to as *input* or *source* images, and the resultant combined image (or images) as *fused* image.

The actual fusion process can take place at different levels of information representation. A common categorization is to distinguish between pixel, feature and symbol level [1], although indeed these levels can be combined themselves [2]. Image fusion at pixel-level means fusion at the lowest processing level referring to the merging of measured physical parameters [3,4]. It generates a fused image in which each pixel is determined from a set of pixels in the various sources. Fusion at feature-level requires first the extraction (e.g., by segmentation procedures) of the features contained in the various input sources [5,6]. Those features can be identified by characteristics such as size, shape, contrast and texture. The fusion is thus based on those extracted features and enables the detection of useful features with higher confidence. Fusion at symbol level allows the information to be effectively combined at the highest level of abstraction [7,8]. The input images are usually processed individually for information extraction and

☆ This work is supported by the Dutch Technology Foundation STW, project no. CWI.4616.
* Tel.: +31-20-592-4214; fax: +31-20-592-4199.
E-mail address: gemma.piella@cwi.nl (G. Piella).
[1] Terminologies such as fusion, integration and merging, are often used interchangeably in the literature.

1566-2535/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S1566-2535(03)00046-0

classification. This results in a number of symbolic representations which are then fused according to decision rules which reinforce common interpretation and resolve differences. The choice of the appropriate level depends on many different factors such as data sources, application and available tools. At the same time, the selection of the fusion level determines the necessary pre-processing involved. For instance, fusing data at pixel-level requires co-registered images at subpixel accuracy because the existing fusion methods are very sensitive to misregistration.

Currently, it seems that most image fusion applications employ pixel-based methods. The advantage of pixel fusion is that the images used contain the original information. Furthermore, the algorithms are rather easy to implement and time efficient. As we observed before, an important pre-processing step in pixel-fusion methods is image registration, which ensures that the information from each source is referring to the same physical structures in the real-world. Throughout this paper, it will be assumed that all source images have been registered. Comprehensive reviews on image registration can be found in [9].

The aim of image fusion is to integrate complementary and redundant information from multiple images to create a composite that contains a 'better' description of the scene than any of the individual source images. This fused image should increase the performance of the subsequent processing tasks. Considering the objectives of image fusion and its potential advantages, some generic requirements can be imposed on the fusion algorithm [10]: (i) it should not discard any salient information contained in the input images; (ii) it should not introduce any artifacts or inconsistencies which can distract or mislead a human observer or any subsequent image processing steps; (iii) it must be reliable, robust and, as much as possible, tolerant of imperfections such as noise or misregistrations.

To illustrate some of the challenges we have to face when developing a fusion algorithm, consider the registered source images in Fig. 1(a) and (b) depicting the same scene. While in the visual image of Fig. 1(a) it is hard to distinguish the person in camouflage from the background, this person is clearly observable in the infrared (IR) image of Fig. 1(b). In contrast, the easily discernible background in the visual image, such as the fence, is nearly imperceptible in the IR image. How to combine both images in a unique composite which represents the overall scene better than any of the two individual images? We sum up explicitly some of the difficulties that we encounter:

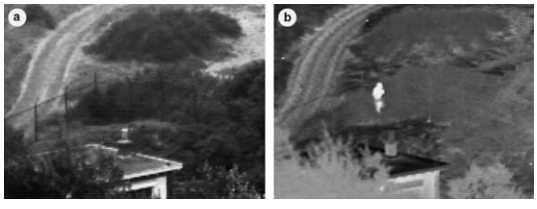

Fig. 1. Example of source images to be fused: (a) visual image; (b) infrared image. Images courtesy of Alex Toet, from TNO Human Factors Institute, The Netherlands.

- Complementary information: some image features appear in one source but not in the other, e.g., the person in Fig. 1(b) or the fence in Fig. 1(a).
- Common but contrast reversal information: there are various objects and regions that occur in both images but with opposite contrast, e.g., part of the roof of the house or the bushes at the left lower corner. Thus, the direct approach of adding and averaging the source images is not satisfactory.
- Disparity between sensors: input images come from different types of sensors which have different dynamic range and different resolution. Moreover, they may not be equally reliable.

Image fusion is widely recognized as a valuable tool for improving overall system performance in image-based application areas such as defense surveillance, remote sensing, medical imaging and computer vision. We list some application fields and give some references to the related literature.

Defense systems. It covers subareas such as detection, identification and tracking of targets [11,12], mine detection [13], tactical situation assessment [14,15], and person authentication [16].

Geoscience. This field concerns the earth study with satellite and aerial images (remote sensing) [1]. The main problem is the interpretation and classification of images. The fused image allows the detection of roads, airports, mountainous areas, etc. [17–19].

Medical imaging. The fusion of multimodal images is very useful for clinical applications such as diagnosis, modeling of the human body or treatment planning [20–22].

Robotics and industrial engineering. Here, fusion is commonly used to identify the environment in which the robot or intelligent system evolves [23,24] and for navigation [25,26]. Image fusion is also employed in industry [27,28].

There are various techniques for image fusion, even at the pixel-level [1]. The selection of the appropriate one depends strongly on the type of application. Some commonly used techniques in pixel-level fusion are: *weighted combination* [29,30]; *optimization approach* [31] and *biologically-based approaches* such as neural networks [32,33] and multiresolution (MR) decompositions [3,34–36].

Henceforth, we confine our discussion to MR image fusion approaches. In particular, we focus on pixel and feature-level MR fusion schemes where the output is a

single fused image which is constructed primarily for display on a computer monitor.

The rest of the report is organized as follows. In Section 2 we review the basics of MR decomposition theory. In Section 3 we present a general framework for pixel-based MR fusion. Within this framework, we describe some of the existing schemes in literature and show fused image examples of existing as well as new fusion schemes. In Section 4 we extend the previous framework and propose a region-based MR fusion strategy. We illustrate different ways of using the region information and present some experimental results using the region approach. In Section 5 we briefly discuss the topic of performance assessment. Finally, in Section 6, we present conclusions and suggest directions for further work.

It is to be noted that the fusion framework in Section 3 has been partially inspired by the MR fusion methodology proposed by Zhang and Blum [37]. The authors proposed also a region-based fusion algorithm [38]. Our region approach, however, is different from theirs in several aspects which will be discussed later in Section 4.

2. Multiresolution decomposition schemes: an overview

A MR decomposition scheme decomposes the signal being analyzed into several components, each of which captures information present at a given scale. MR methods in signal and image processing are very important for various reasons: (i) real-world objects usually consist of structures at different scales; (ii) there is strong evidence that the human visual system (HVS) processes information in a MR fashion; (iii) MR methods offer computational advantages and, moreover, appear to be robust. In the following, MR decompositions are described within the axiomatic framework of Heijmans and Goutsias [39,40].

2.1. Decomposition systems with perfect reconstruction

The idea of a decomposition system with perfect reconstruction is to obtain a more convenient representation (*analysis*) of the signal such that no information is lost, i.e., the signal can be recovered through some reconstruction process (*synthesis*). Fig. 2 depicts a general scheme for the decomposition of an input signal $x^{(0)} \in V_0$ into two components $(x^{(1)}, y^{(1)}) \in V_1 \times W_1$. Here, $x^{(1)}$ and $y^{(1)}$ can be interpreted as the *approximation* and *detail* signals of $x^{(0)}$, respectively. In other words, $x^{(1)}$ is a sort of 'simplification' of $x^{(0)}$, inheriting many of its properties, whereas $y^{(1)}$ is a kind of 'refinement' that contains the information that has been discarded in the simplification process. The operators $\psi^{\uparrow} : V_0 \mapsto V_1$, $\omega^{\uparrow} : V_0 \mapsto W_1$ are called *analysis operators* and the operator $\Psi^{\downarrow} : V_1 \times W_1 \mapsto V_0$ is called the *synthesis operator*. The

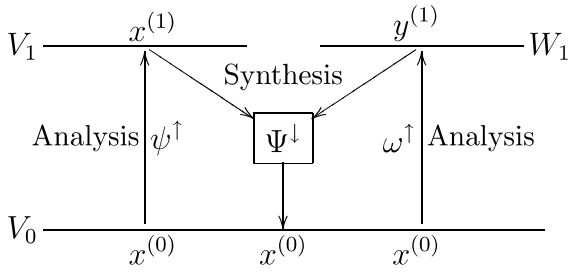

Fig. 2. A signal decomposition scheme with perfect reconstruction.

assumption that no information is lost by the decomposition is expressed by the requirement that $\Psi^{\downarrow}$ is the left inverse of $\Psi^{\uparrow} = (\psi^{\uparrow}, \omega^{\uparrow})$, i.e., $\Psi^{\downarrow}(\psi^{\uparrow}(x^{(0)}), \omega^{\uparrow}(x^{(0)})) = x^{(0)}$, for $x^{(0)} \in V_0$. This condition is referred to as the *perfect reconstruction condition*.

In various signal and image applications, the decomposition $x^{(0)} \mapsto (x^{(1)}, y^{(1)})$ is only a first step toward an analysis of $x^{(0)}$. Subsequent steps comprise a decomposition of $x^{(1)}$ into $x^{(2)}$ and $y^{(2)}$, of $x^{(2)}$ into $x^{(3)}$ and $y^{(3)}$, and so forth. By concatenating several systems of the form depicted in Fig. 2 we obtain a *multilevel decomposition system*. If the higher levels are obtained by means of some spatial filtering (e.g., linear or morphological) of the lower level signals, possibly followed by a sampling step, then we call the system a *multiresolution* or *multiscale* decomposition scheme.

To formalize this procedure, assume that there exists a sequence of signal spaces V_k, $k \geqslant 0$, and detail spaces W_k, $k \geqslant 1$. At each level $k \geq 0$ we have two analysis operators, $\psi_k^{\uparrow} : V_k \mapsto V_{k+1}$ and $\omega_k^{\uparrow} : V_k \mapsto W_{k+1}$, and a synthesis operator $\Psi_k^{\downarrow} : V_{k+1} \times W_{k+1} \mapsto V_k$, satisfying the perfect reconstruction condition:

$$\Psi_k^{\downarrow}(\psi_k^{\uparrow}(x), \omega_k^{\uparrow}(x)) = x, \quad \text{for } x \in V_k. \tag{1}$$

A given input signal $x^{(0)} \in V_0$ can be decomposed by the recursive scheme

$$\begin{aligned} x^{(0)} &\to \{y^{(1)}, x^{(1)}, \} \to \{y^{(1)}, y^{(2)}, x^{(2)}\} \to \cdots \\ &\to \{y^{(1)}, \ldots, y^{(K-1)}, y^{(K)}, x^{(K)}\}, \end{aligned} \tag{2}$$

where

$$\begin{cases} x^{(k+1)} = \psi_k^{\uparrow}(x^{(k)}) \\ y^{(k+1)} = \omega_k^{\uparrow}(x^{(k)}) \end{cases} \quad k = 0, 1, \ldots, K-1.$$

Here, $x^{(k+1)}$ is an approximation of $x^{(k)}$, but can also be regarded as a '$(k+1)$st-order' coarse approximation of the original signal $x^{(0)}$. In contrast, the detail signal $y^{(k+1)}$ contains information about $x^{(k)}$ that is not present in the simplified component $x^{(k+1)}$. Note that, because of the perfect reconstruction condition, the original signal $x^{(0)}$ can be perfectly reconstructed from $x^{(K)}$ and $y^{(1)}, y^{(2)}, \ldots, y^{(K)}$ by means of the backward recursion:

$$x^{(k)} = \Psi_k^{\downarrow}(x^{(k+1)}, y^{(k+1)}), \quad k = K-1, K-2, \ldots, 0. \tag{3}$$

Fig. 3 illustrates the analysis and synthesis schemes for the particular case where $K = 3$.

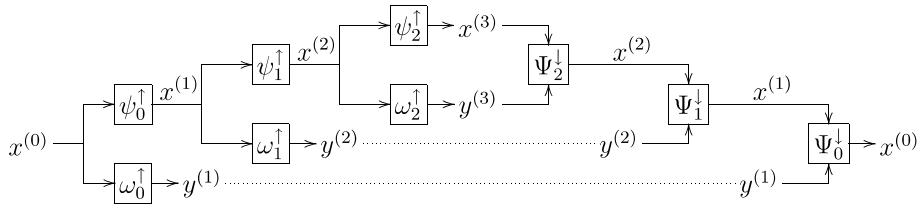

Fig. 3. A three-level decomposition system: analysis and synthesis.

2.2. The pyramid transform

The pyramid transform is characterized by the assumption that

$$\Psi_k^{\downarrow}(x,y) = \psi_k^{\downarrow}(x) + y, \quad \text{for } x \in V_{k+1},\ y \in W_{k+1}, \tag{4}$$

where $W_{k+1} \subseteq V_k$ and $\psi_k^{\downarrow} : V_{k+1} \mapsto V_k$. The perfect reconstruction condition in (1) can be reformulated as

$$\psi_k^{\downarrow}\psi_k^{\uparrow}(x) + \omega_k^{\uparrow}(x) = x, \quad \text{for } x \in V_k.$$

Thus, $\omega_k^{\uparrow}(x) = x - \psi_k^{\downarrow}\psi_k^{\uparrow}(x)$ is the error of the synthesis operator $\psi_k^{\downarrow}$ when reconstructing x from the approximation $\psi_k^{\uparrow}(x)$. In this case, the recursive analysis scheme in (2) is given by

$$\begin{cases} x^{(k+1)} = \psi_k^{\uparrow}(x^{(k)}) \\ y^{(k+1)} = x^{(k)} - \psi_k^{\downarrow}(x^{(k+1)}) \end{cases} \quad k = 0, 1, \ldots, K-1, \tag{5}$$

and the synthesis step in (3) is

$$x^{(k)} = \psi_k^{\downarrow}(x^{(k+1)}) + y^{(k+1)}, \quad k = K-1, K-2, \ldots, 0. \tag{6}$$

We refer to the decomposition process $x^{(0)} \mapsto \{y^{(1)}, \ldots, y^{(K-1)}, y^{(K)}, x^{(K)}\}$ by means of (5) as the *pyramid transform* of $x^{(0)}$, and to the process of synthesizing $x^{(0)}$ by means of (6) as the *inverse pyramid transform*. A block diagram illustrating the pyramid transform and its inverse is shown in Fig. 4. We call the sequence $\{x^{(0)}, x^{(1)}, \ldots, x^{(K)}\}$ the *approximation pyramid* and the sequence $\{y^{(1)}, y^{(2)}, \ldots, y^{(K)}\}$ the *detail pyramid*.

The axiomatic pyramid approach described above encompasses several existing pyramid techniques such as the well-known Laplacian pyramid introduced by Burt and Adelson [41]. Obviously, the choice of different analysis and synthesis operators results in different kinds of pyramids. In particular, non-linear pyramids have attracted a great deal of attention. A well-known example is the morphological pyramid [39,42], where the analysis and synthesis operators are based on morphological operators [43]. Another instance of non-linear pyramids is the ratio-of-low-pass pyramid [44]. Here, the ratio (rather than the standard difference) of the successive low-pass filtered signals is computed. In fact, the operator '+' used in (4) can be replaced by any invertible operation (see [39] for details).

Observe that a signal representation obtained by means of a pyramid transform (i.e., detail signals along with the coarsest approximation) is overcomplete in the sense that it produces more samples than the original signal. This is a direct consequence of the fact that the detail signal $y^{(k)}$ 'lives' at the same resolution [2] level as $x^{(k-1)}$.

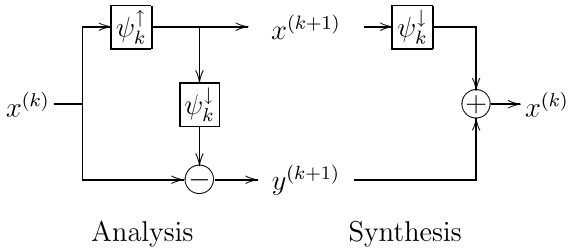

Fig. 4. Pyramid transform (analysis) and its inverse (synthesis).

2.3. Other multiresolution decompositions

2.3.1. The wavelet transform

A general wavelet decomposition has the structure depicted in Fig. 2, but in addition to the perfect reconstruction condition (1), i.e.,

$$\Psi^{\downarrow}(\psi^{\uparrow}(x), \omega^{\uparrow}(x)) = x, \quad \text{for } x \in V_0, \tag{7}$$

it satisfies the additional constraints

$$\psi^{\uparrow}(\Psi^{\downarrow}(x,y)) = x \quad \text{and} \quad \omega^{\uparrow}(\Psi^{\downarrow}(x,y)) = y, \quad \text{for } x \in V_1,\ y \in W_1, \tag{8}$$

which guarantee that the decomposition is non-redundant. Note that (7) and (8) imply that the analysis operator $\Psi^{\uparrow} = (\psi^{\uparrow}, \omega^{\uparrow})$ and the synthesis operator $\Psi^{\downarrow}$ are inverses of each other. Concatenation of a series of analysis steps yields a MR decomposition called the *wavelet transform*. Often, e.g., in the linear case, the synthesis operator $\Psi^{\downarrow}$ is of the special form

$$\Psi^{\downarrow}(x,y) = \psi^{\downarrow}(x) + \omega^{\downarrow}(y), \quad x \in V_1,\ y \in W_1. \tag{9}$$

Fig. 5 diagrams the corresponding synthesis part of such a wavelet decomposition for the multilevel case where $K = 3$.

It was explained by Heijmans and Goutsias [40] how existing linear wavelets [45,46] can fit into this abstract wavelet scheme. Note, however, that the expressions in

[2] Here, the term 'resolution' refers to the size of the signal.

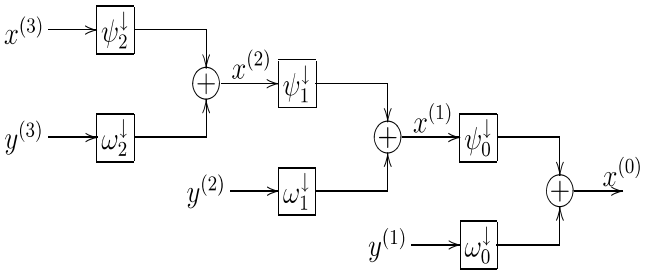

Fig. 5. Synthesis scheme of a three-level uncoupled wavelet decomposition system.

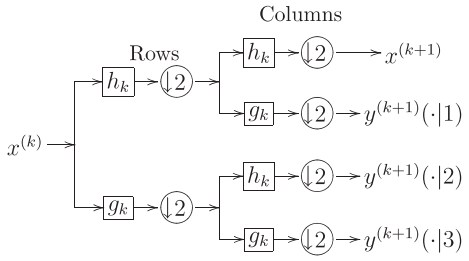

Fig. 6. Separable two-dimensional wavelet transform.

(7) and (8) are formulated in operator terms, and do not require any sort of linearity assumption or inner product. This allows a broad class of non-linear wavelet decomposition schemes.

One drawback of the discrete wavelet transform (DWT) and, to a lesser extent, also for the pyramid transform, is that it generally yields a shift-variant signal representation. This means that a simple shift of the input signal may lead to complete different transform coefficients. The lack of translation invariance can be avoided if the outputs of the filter banks are not decimated. The resulting undecimated wavelet transform [46] yields a redundant MR representation where the approximation and detail signals have all the same size as the original signal.

Most wavelet and pyramid transforms have been designed in the one-dimensional case. By successive application of such one-dimensional transforms on the rows and the columns (or vice versa) of an image, one obtains a so-called *separable* two-dimensional transform. This construction is illustrated in Fig. 6 for the wavelet transform.

At each level k, the input $x^{(k)}$ is decomposed into a coarse approximation $x^{(k+1)}$ and three detail signals $y^{(k+1)} = \{y^{(k+1)}(\cdot|1), y^{(k+1)}(\cdot|2), y^{(k+1)}(\cdot|3)\}$, corresponding to the horizontal, vertical and diagonal directions. Non-separable transforms [40,47] provide decompositions with more general properties but they have been used less often in image applications due to the lack of general tools for their design.

2.3.2. Wavelet packets

With a straightforward generalization of the wavelet transform, we can obtain an even sparser representation of a signal. Instead of dividing only the approximation spaces V_k to derive the approximation and detail spaces V_{k+1} and W_{k+1}, we divide the detail spaces as well. The recursive splitting of spaces can be represented in a binary tree as illustrated in Fig. 7. At each node of the tree, we have the option to split or not. This allows the construction of an arbitrary dyadic tree structure. Each structure is associated with a function basis known as a *wavelet packet basis* [46]. They generalize the fixed dyadic construction of the standard wavelet basis by decomposing the frequency axis in intervals of varying sizes. Given a signal (or class of signals) and a fixed set of filters, we can obtain the 'best' (according to some criterion) tree decomposition.

2.3.3. Local basis

A local basis divides the time axis into intervals of varying sizes. Of particular interest are the cosine bases [46], which are obtained by designing smooth windows that cover each time interval and multiplying them by cosine functions of different frequencies.

Similarly to wavelet packets, a local cosine tree can be constructed by recursively dividing spaces built with local cosine bases. As in the wavelet packet case, this offers the possibility of choosing a 'best' basis for a given signal. A best local cosine basis adapts the time segmentation to the variations of the signal time-frequency structures. In comparison with wavelet packets, we gain time adaptation but we lose frequency flexibility (since the frequency axis is being split with constant bandwidth).

2.3.4. Multiwavelets

Multiwavelet decompositions offer more design flexibility by introducing (at each level) several analysis and

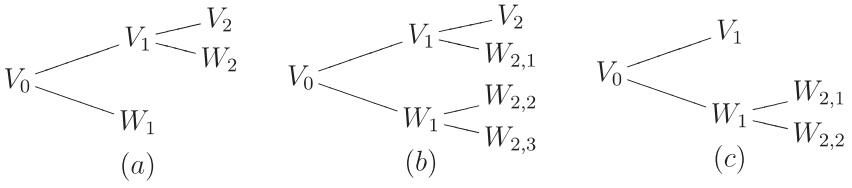

Fig. 7. Examples of tree-structure filter banks of depth 2: (a) only low-pass is split (standard wavelet); (b) full tree; (c) only high-pass is split.

synthesis operators. Multiwavelets have some advantages over scalar wavelets in relation to properties which are known to be important in signal processing such as short support, orthogonality, symmetry, and vanishing moments. A scalar wavelet, except for the Haar system, cannot possess all these properties at the same time. In contrast, a multiwavelet system can simultaneously provide perfect reconstruction, orthogonality, linear-phase symmetry and a high order of approximation (vanishing moments) [48]. The main drawback, however, is that they are implemented with more complicated filter banks than the standard wavelet transforms.

2.3.5. Steerable pyramid

The steerable pyramid is an overcomplete, linear, multiresolution and multiorientation image decomposition where the analysis and synthesis operators are (in the simplest case) derivative operators with different supports and orientations. The associated filters are such that the resulting transform is self-inverting (i.e., the synthesis filters are just a reflected version of the analysis filters) and, moreover, it is translation and rotation-invariant. The system diagram for the steerable pyramid (both analysis and synthesis) is depicted in Fig. 8.

Initially, the image is separated by the pre-processing filters H_0 and G_0 into a low and a high-pass subbands. We denote the high-pass signal by $z^{(1)}$. The low-pass branch is then divided into a set of P oriented detail images and one approximation image. The detail images $y^{(1)}(\cdot|p)$, $p=1,\ldots,P$, are obtained using the band-pass filters $B_1,\ldots,B_P$; while the approximation signal $x^{(1)}$ is obtained using a low-pass filter H_1 followed by a dyadic downsampling. The process of splitting into P details and one approximation is iterated on the approximation image (thus, filters H_0 and G_0 are not used in the successive levels).

In order to ensure that the transform is invertible as well as jointly invariant in orientation and position, the filters must satisfy specific radial (scale) and angular (orientation) frequency constraints [49]. The pyramid can be designed to produce any number of orientation bands P, resulting in an overcomplete transform by a factor of $4P/3$.

2.3.6. Gradient pyramid

A gradient pyramid [50] is obtained by applying a gradient operator to each level of the Gaussian pyramid $\{x^{(k)}\}$, $k=0,\ldots,K$. Each image $x^{(k)}$ is filtered by a set of four oriented gradient filters g_p, $p=1,\ldots,4$. The resulting filtered subbands correspond to the detail images $y^{(k+1)}(\cdot|p)$, $p=1,\ldots,4$, representing the horizontal, vertical and the two diagonals directions. To reconstruct the original image from this gradient decomposition, a Laplacian pyramid is constructed as intermediate result. First, a (derivative) synthesis filter $\tilde{g}_p$ is applied to $y^{(k+1)}(\cdot|p)$, $p=1,\ldots,4$. A Laplacian pyramid $\{y_{\mathrm{L}}^{(k+1)}\}$, can then be obtained by summing up, at each level, the filtered resulting images. Fig. 9 illustrates one level of the gradient pyramid transform. Here, h is the low-pass filter used to construct the Gaussian pyramid $\{x^{(k)}\}$, $k=0,\ldots,K$, and $\tilde{h}$ its corresponding synthesis filter.

2.4. Notation

Henceforth, the MR decomposition of an image $x^{(0)}$ is denoted by y and it is assumed to be of the form

$$y=\{y^{(1)},y^{(2)},\ldots,y^{(K)},x^{(K)}\}. \tag{10}$$

Here $x^{(K)}$ represents the approximation image at the highest level (lowest resolution) of the MR structure, while images $y^{(k)}$, $k=1,\ldots,K$, represent the detail images at level k. The detail at level k will, in general, comprise various frequency or orientation bands, depending on the type of MR transform that has been used. We assume henceforth that $y^{(k)}$ is composed of P detail images, i.e., $y^{(k)}=\{y^{(k)}(\cdot|1),\ldots,y^{(k)}(\cdot|P)\}$.

Let $I_x^{(k)}$ and $I_y^{(k)}(p)$ denote the domain of $x^{(k)}$ and $y^{(k)}(\cdot|p)$, respectively. We use the vector coordinate $\boldsymbol{n}=(n,m)$ to index the location of the coefficient. Then $x^{(k)}(\boldsymbol{n})$, where $\boldsymbol{n}\in I_x^{(k)}$, represents the approximation coefficient at location $\boldsymbol{n}$ within level k. Similarly, $y^{(k)}(\boldsymbol{n}|p)$, where $\boldsymbol{n}\in I_y^{(k)}(p)$, represents the detail coefficient at lo-

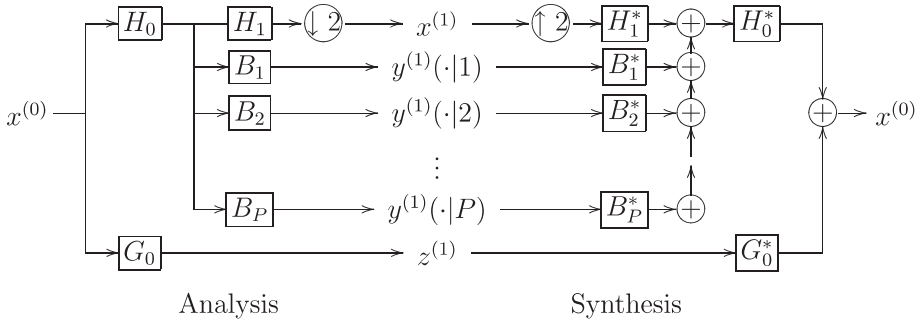

Fig. 8. Steerable transform (analysis) and its inverse (synthesis).

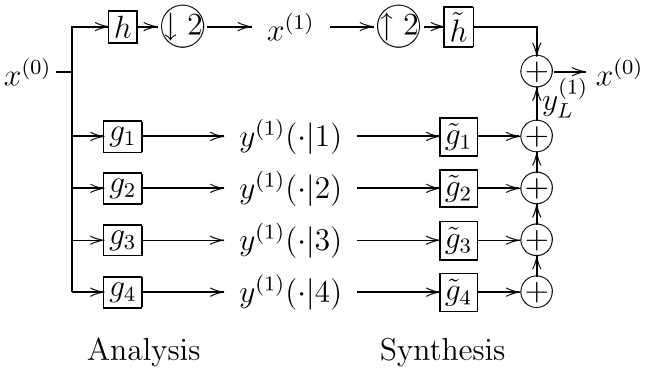

Fig. 9. Gradient pyramid transform (analysis) and its inverse (synthesis).

cation $\boldsymbol{n}$ within level k and band p. Note that $I_x^{(k)}$ is not necessarily equal to $I_y^{(k)}(p)$. In the pyramid case, for example, $y^{(k)}$ represents a detail image of the same size as $x^{(k-1)}$, while in the standard wavelet transform $y^{(k)}(\cdot|p)$ is a detail image of the same dimensions as $x^{(k)}$. Note also that in most cases $I_y^{(k)}(p)$ does not depend on p.

For convenience, we will sometimes denote the approximation image $x^{(k)}$ by $y^{(k)}(\cdot|0)$. In this way, we can use the general expression of $y^{(k)}(\cdot|p)$ to refer both to the detail images (for $p = 1, \ldots, P$) and the approximation image (for $p = 0$). If no confusion is possible, we will use the shorthand notation $(\cdot)$ to denote $(\boldsymbol{n}|p)$; e.g., we will write $y^{(k)}(\cdot)$ rather than $y^{(k)}(\boldsymbol{n}|p)$.

3. The general pixel-based MR fusion scheme

The basic idea underlying the MR-based image fusion approach is to perform a MR transform on each source image and, following some specific fusion rules, construct a composite MR representation from these inputs. The fused image is obtained by applying the inverse transform on this composite MR representation. This process is illustrated in Fig. 10 for the case of two input source images.

In the literature one finds several variants of the MR fusion scheme. In what follows, we present a general framework which encompasses most of them. Section 3.1 describes the various modules the framework consists of. Within the framework, some of the existing algorithms proposed in literature are reviewed in Section 3.2. Examples of such schemes as well as other implementation alternatives are given in Section 3.3.

3.1. The general framework

In Fig. 11 we show a more detailed version of the fusion scheme of Fig. 10, in which the combination algorithm has been specified. In our framework, the combination algorithm consists of four modules: the *activity* and *match* measures extract information from the MR decompositions of the inputs, which is then used by the *decision* and *combination* map to compute the MR decomposition of the fused image.

Below, we give a short description of each of the building blocks. Note, however, that some of them, such as the 'match block' are optional.

MR analysis (Ψ): This block computes a MR decomposition of the input sources x_S, $S \in \mathscr{S}$, where $\mathscr{S}$ is the index set of source images. For every input x_S we obtain its MR representation $y_S = \Psi(x_S)$, with y_S having the form defined in (10).

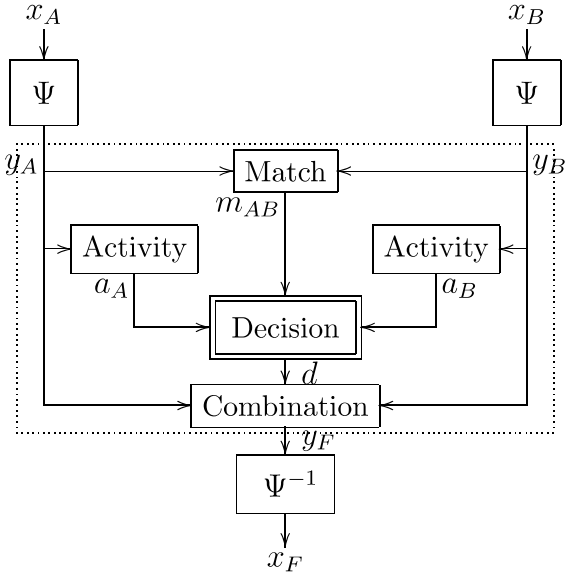

Fig. 11. Generic pixel-based MR fusion scheme.

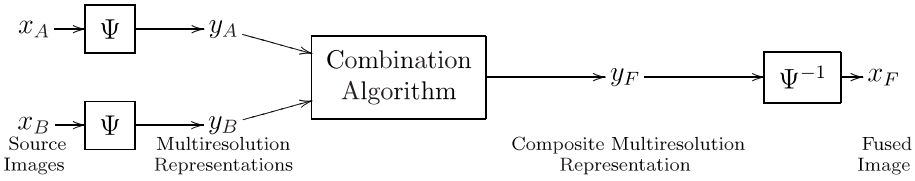

Fig. 10. Multiresolution image fusion scheme.

Activity measure: The degree to which each coefficient in y_S is salient (i.e., of interest for a task at hand) will be expressed by the so-called *activity*. The activity function block associates to every band image $y_S^{(k)}(\cdot|p)$ an activity $a_S^{(k)}(\cdot|p)$, which reflects the local activity of the image.

Match measure: This measure is supposed to quantify the degree of 'similarity' between the sources. More precisely, the match value $m_{AB}^{(k)}(\cdot)$ reflects the resemblance between the inputs $y_A^{(k)}(\cdot)$ and $y_B^{(k)}(\cdot)$.

Decision map: This block is the core of the combination algorithm. Its output governs the actual combination of the coefficients of the MR decompositions of the various sources. For each level k, orientation band p, and location $\boldsymbol{n}$, the decision process assigns a value $\delta = d^{(k)}(\boldsymbol{n}|p)$ which is then used for the computation of the composite $y_F^{(k)}(\boldsymbol{n}|p)$.

Combination map: This module describes the actual combination of the transform coefficients of the sources. For each level k, orientation band p, and location $\boldsymbol{n}$, the combination map yields the composite coefficient $y_F^{(k)}(\boldsymbol{n}|p)$.

MR synthesis (Ψ^{-1}): Finally, the fused image is obtained by applying the inverse transformation on the composite MR decomposition y_F, that is, $x_F = \Psi^{-1}(y_F)$, where Ψ^{-1} is the inverse MR transform.

From the previous description, one can see that the parameters and functions comprised by the different blocks can be chosen in several ways. In the following, we discuss them in more detail.

3.1.1. MR analysis and synthesis

As we have seen in Section 2, the MR representation y_S comprises information at different scales. High-levels contain coarse scale information while low-levels contain finer details. Such a representation is suitable for image fusion, not only because it enables one to consider and fuse image features separately at different scales, but also because it produces large coefficients near edges, thus revealing salient information [51].

Basically, the issues to be addressed are the specific type of MR decomposition (pyramid, wavelet, linear, morphological, etc.) and the number of decomposition levels.

A large part of research on MR image fusion has focused on choosing an appropriate MR representation which facilitates the selection and combination of salient features. Studying the existing literature, we draw the following conclusions:

- In general, sampling causes a deterioration in the quality of the fused image by introducing heavier blocking effects than would have obtained by using decompositions without sampling.
- Shift and rotation-invariance properties are often required. For many applications, the fusion result should not depend on the location or orientation of the objects in the input sources. Shift and rotation dependency are especially undesirable considering misregistration problems or for image sequence fusion.
- In linear approaches, the specific filter used has little influence on the fusion result; shorter filters lead to slightly sharper fusion results.
- MR decompositions constructed with morphological techniques are more suited for the analysis of shape and size of specific features in the images.

Another parameter which influences performance is the number of decomposition levels (analysis depth). To perform a consistent fusion of objects at arbitrary scales, the decomposition over a large number of scales may appear necessary. However, using more levels does not necessarily produce better results; it may produce low-resolution bands where neighboring features overlap. This gives rise to discontinuities in the composite representation and, thus, introduces distortions, such as blocking effects or 'ringing' artifacts, into the fused image. The required analysis depth is primarily related to the spatial extent of the relevant objects in the source images. In general, it is not possible to compute the optimal analysis depth, but as a rule of thumb, the larger the objects of interest are, the higher the number of decomposition levels should be.

3.1.2. Activity measure

The meaning of 'saliency' (and thus the computation of the activity) depends on the nature of the source images as well as on the particular fusion application. Generally, based on the fact that the HVS is primarily sensitive to local contrast changes (i.e., edges), most fusion algorithms compute the activity as some sort of energy calculation, e.g.,

$$a_S^{(k)}(\boldsymbol{n}|p) = \sum_{\Delta\boldsymbol{n}\in\mathscr{W}^{(k)}(p)} w^{(k)}(\Delta\boldsymbol{n}|p)|y_S^{(k)}(\boldsymbol{n}+\Delta\boldsymbol{n}|p)|^{\gamma}, \qquad \gamma\in\mathbb{R}_+, \tag{11}$$

where $\mathscr{W}^{(k)}(p)$ is a finite window at level k and orientation p, and $w^{(k)}(\cdot|p)$ are the window's weights. In the simplest case, the activity is just the absolute value of the coefficient, that is,

$$a_S^{(k)}(\cdot) = |y_S^{(k)}(\cdot)|. \tag{12}$$

Alternatively, the contrast of the component with its neighbors, or some other linear or non-linear criteria can provide that measure.

In practice, the window $\mathscr{W}^{(k)}(p)$ over which the function operates is small, typically including only the sample itself (*sample-based* operation), or a 3×3, or 5×5 window centered at the sample (*area-based* operation). However, other size and shape templates have also been used. Increasing the size of the neighborhood from the simple sample-based case, adds robustness to the fusion system as it provides a smooth activity function. How-

ever, larger templates cause problems at lower resolution levels when their size exceeds the size of the most salient features.

3.1.3. Match measure

The match or similarity between the transform coefficients of the source images is usually expressed in terms of a local correlation measure. Alternatively, the relative amplitude of the coefficients or some other criteria can be used. In the following expression, the match value between $y_A^{(k)}(\cdot)$ and $y_B^{(k)}(\cdot)$ is defined as a normalized correlation averaged over a neighborhood of the samples:

$$m_{AB}^{(k)}(\boldsymbol{n}|p) = \frac{2\sum_{\Delta\boldsymbol{n}\in\mathscr{W}^{(k)}(p)} w^{(k)}(\Delta\boldsymbol{n}|p)\, y_A^{(k)}(\boldsymbol{n}+\Delta\boldsymbol{n}|p)\, y_B^{(k)}(\boldsymbol{n}+\Delta\boldsymbol{n}|p)}{\sum_{\Delta\boldsymbol{n}\in\mathscr{W}^{(k)}(p)} w^{(k)}(\Delta\boldsymbol{n}|p)\left(|y_A^{(k)}(\boldsymbol{n}+\Delta\boldsymbol{n}|p)|^2 + |y_B^{(k)}(\boldsymbol{n}+\Delta\boldsymbol{n}|p)|^2\right)},$$

where $\mathscr{W}^{(k)}(p)$ is the window at level k and orientation p, and $w^{(k)}(\cdot|p)$ its corresponding weights. By analyzing the match measure, one can determine where the sources differ and to which extent, and use this information to combine them in an appropriate way.

3.1.4. Combination map

This module performs the actual combination of the MR coefficients of the sources. For simplicity, consider two sources and assume that every composite coefficient is 'assembled' from the source coefficients at the corresponding level, band and position. More precisely,

$$y_F^{(k)}(\cdot) = C^{(k)}(y_A^{(k)}(\cdot),\ y_B^{(k)}(\cdot),\ d^{(k)}(\cdot)),$$

where $C^{(k)} : \mathbb{R}^3 \mapsto \mathbb{R}$ is the combination map at level k.

A simple choice for $C^{(k)}$ is a linear mapping, e.g.,

$$C^{(k)}(y_1, y_2, \delta) = w_A(\delta)y_1 + w_B(\delta)y_2, \tag{13}$$

where the weights $w_A(\delta)$, $w_B(\delta)$ depend on the decision parameter δ.

Although non-linear mappings [52,53] are another option, we restrict ourselves to linear combination maps as in (13), yet with possibly more than two input sources. Thus, the composite coefficients $y_F^{(k)}(\cdot)$ are obtained by an *additive or weighted combination*:

$$y_F^{(k)}(\cdot) = \sum_{S\in\mathscr{S}} w_S(d^{(k)}(\cdot))\, y_S^{(k)}(\cdot). \tag{14}$$

For the particular case where only one of the coefficients $y_S^{(k)}(\cdot)$ has a weight distinct from zero, that is, only one of the sources contributes to the composite, we talk about *selective combination* or *combination by selection*.

3.1.5. Decision map

The construction of the decision map is a key point because its output $d^{(k)}$ governs the combination map $C^{(k)}$. Therefore, the decision map actually determines the combination of the various MR decompositions y_S and, hence, the construction of the composite y_F.

In our case, where we assume a weighted combination such as in (14), the decision map controls the values of the weights to be assigned to each of the source coefficients. Indeed, specifying the decision $\delta = d^{(k)}(\cdot)$ is, in practice, equivalent to specifying the weights $w_S(\delta)$. For this reason, the combination and decision maps are often 'grouped' together by expressing the composite coefficients in terms of the parameters or functions the decision is based on. The problem of 'how to compute $d^{(k)}(\cdot)$' is translated into the problem of 'how to compute $w_S(d^{(k)}(\cdot))$'. A natural approach is to assign to each coefficient a weight that depends increasingly on the activity. In general, the resulting weighted average (performed by the combination map) leads to a stabilization of the fusion result, but it introduces the problem of contrast reduction in case of opposite contrast in different source images. This can be avoided by using a selective rule where the most salient component, i.e., the one with largest activity, is chosen for the composite. In this case, after the combination map we get

$$y_F^{(k)}(\cdot) = y_M^{(k)}(\cdot) \quad \text{with } M = \arg\max_{S\in\mathscr{S}}(a_S^{(k)}(\cdot)). \tag{15}$$

In other words, the decision process 'decides' that the most salient coefficient (among the various $y_S^{(k)}(\cdot)$, $S \in \mathscr{S}$) is the best choice for the composite coefficient $y_F^{(k)}(\cdot)$, and 'tells' the combination process to select it, i.e.,

$$w_S(d^{(k)}(\cdot)) = \begin{cases} 1 & \text{if } S = \arg\max_{S'\in\mathscr{S}}(a_{S'}^{(k)}(\cdot)), \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

This selective combination is also known in the literature as a 'choose max' or *maximum selection* rule. It works well under the assumption that at each image location, only one of the source images provides the most useful information. This assumption is not always valid, and a weighted combination may appear a better option. Alternatively, a match measure can be used to decide how to combine the coefficients. For instance,

$$y_F^{(k)}(\cdot) = \begin{cases} y_A^{(k)}(\cdot) & \text{if } m_{AB}^{(k)}(\cdot) \leqslant T \text{ and } a_A^{(k)}(\cdot) > a_B^{(k)}(\cdot), \\ y_B^{(k)}(\cdot) & \text{if } m_{AB}^{(k)}(\cdot) \leqslant T \text{ and } a_A^{(k)}(\cdot) \leqslant a_B^{(k)}(\cdot), \\ \frac{y_A^{(k)}(\cdot)+y_B^{(k)}(\cdot)}{2} & \text{otherwise;} \end{cases}$$

for some threshold T. Thus, at sample locations where the source images are distinctly different, the combination process selects the most salient component, while at sample locations where they are similar, the process averages the source components. In this manner, averaging

reduces noise and provides stability where source images contain similar information, whereas selection retains salient information and reduces artifacts due to opposite contrast.

In the examples presented so far, the decision is taken for each coefficient, without reference to the others. This may degrade the fusion result since there is the possibility of feature cancellation when the inverse transform is applied to obtain the fused image. Taking into account the spatial, inter- and intra-scale dependencies between the coefficients may provide a partial solution to this problem. Note that by construction, each coefficient of a MR decomposition has a set of 'family-related' components in other orientation bands and other levels: they represent the same (or nearby) spatial location in the original image. It seems reasonable then to consider all (or a set of) these coefficients when determining the composite MR representation. A simple example is

$$y_{\mathrm{F}}^{(k)}(\boldsymbol{n}|p) = \begin{cases} y_{\mathrm{A}}^{(k)}(\boldsymbol{n}|p) & \text{if } \sum_{p'=1}^{p'=P} a_{\mathrm{A}}^{(k)}(\boldsymbol{n}|p') > \sum_{p'=1}^{p'=P} a_{\mathrm{B}}^{(k)}(\boldsymbol{n}|p'), \\ y_{\mathrm{B}}^{(k)}(\boldsymbol{n}|p) & \text{otherwise,} \end{cases}$$

where intra-scale dependencies are used. In this particular example, the decision is made globally for a group of samples: for all bands p, all samples in the same level k and location $\boldsymbol{n}$ are assigned the same decision.

Another possibility is to exploit spatial redundancy between neighboring samples. One may assume that spatially close samples are likely to belong to the same image feature and, thus, should be treated in the same way. An illustrative example is the consistency verification method proposed by Li et al. [35]. This method consists in applying a majority filter to a preliminary decision map $\tilde{d}^{(k)}$. Now the filtered decision map determines the combination of the images $y_S^{(k)}$. For example, if according to the preliminary decision $\tilde{d}^{(k)}(\cdot)$, the composite $y_{\mathrm{F}}^{(k)}(\cdot)$ should come from $y_{\mathrm{A}}^{(k)}$, while the majority of the surrounding composite coefficients should come from $y_{\mathrm{B}}^{(k)}$, the decision $\tilde{d}^{(k)}(\cdot)$ is changed so that the composite $y_{\mathrm{F}}^{(k)}(\cdot)$ comes from $y_{\mathrm{B}}^{(k)}$.

This kind of decision methods attempt to exploit the fact that significant image features tend to be stable with respect to variations in space, scale and orientation. Thus, when comparing the corresponding image features in multiple source images, considering the dependencies between the transform coefficients may lead to a more robust fusion strategy.

3.1.6. Combination of approximation images versus combination of detail images

Because of their different physical meaning, the approximation and detail images are usually treated by the combination algorithm through different procedures. Detail coefficients having large absolute values correspond to sharp intensity changes and hence to salient features in the image such as edges, lines and region boundaries. The nature of the approximation coefficients, however, is different. The approximation image $y_S^{(K)}(\cdot|0)$ is a coarse representation of the original image x_S and may have inherited some of its properties such as the mean intensity or texture information. Thus, coefficients $y_S^{(K)}(\boldsymbol{n}|0)$ with high magnitudes do not necessarily correspond with salient features. In this case, an activity measure $a_S^{(K)}(\cdot|0)$ based, for example, on entropy, variance or texture criteria, may be a better alternative than one based on energy like in (11).

In many approaches, the composite approximation coefficients of the highest decomposition level, representing the mean intensity, are taken to be a weighted average of the approximation of the sources:

$$y_{\mathrm{F}}^{(K)}(\boldsymbol{n}|0) = \frac{\sum_{S\in\mathscr{S}} y_S^{(K)}(\boldsymbol{n}|0)}{|S|}, \tag{17}$$

where $|S|$ is the number of sources. The logic behind this combination relies on the assumptions that the sources x_S are contaminated by additive Gaussian noise and that, provided that K is high enough, the relevant features have already been captured by the details $y_S^{(k)}(\cdot|p)$. Thus, the approximation images $y_S^{(K)}(\cdot|0)$ of the various sources contain mostly noise and averaging them reduces the variance of the noise while ensuring that an appropriate mean intensity is maintained.

A popular way to construct the composite y_{F} is to use (17) for the approximation coefficients and the selective combination in (15) for the detail coefficients. For the simple case where $a_S^{(k)}(\cdot) = |y_S^{(k)}(\cdot)|$, and there are two input sources, we can express the combination algorithm as

$$y_{\mathrm{F}}^{(K)}(\boldsymbol{n}|0) = \frac{y_{\mathrm{A}}^{(K)}(\boldsymbol{n}|0) + y_{\mathrm{B}}^{(K)}(\boldsymbol{n}|0)}{2}, \tag{18}$$

$$y_{\mathrm{F}}^{(k)}(\boldsymbol{n}|p) = \begin{cases} y_{\mathrm{A}}^{(k)}(\boldsymbol{n}|p) & \text{if } |y_{\mathrm{A}}^{(k)}(\boldsymbol{n}|p)| > |y_{\mathrm{B}}^{(k)}(\boldsymbol{n}|p)|, \\ y_{\mathrm{B}}^{(k)}(\boldsymbol{n}|p) & \text{otherwise.} \end{cases} \quad p = 1, \ldots, P. \tag{19}$$

Note that other factors may be incorporated for the fusion rules. In particular, if some prior knowledge is available, all the fusion blocks can use such information to improve fusion performance. For instance, when combining the source coefficients, the weights assigned to them may depend not only on the activity and match measure, but may also reflect some a priori knowledge of a specific type, giving preference to certain levels k, spatial positions $\boldsymbol{n}$, or some input sources.

Finally, we want to remark that the decision on which techniques to use is very much driven by the application. At the same time, the characteristics of the resultant fused image depend strongly on the applied

pre-processing and the chosen fusion techniques. The different options we have presented are neither exhaustive nor mutually exclusive and they should merely be considered as practically important examples.

3.2. Overview of some existing fusion schemes

In the literature one finds several MR fusion approaches which fit into our general scheme. In this section, we review some of them. The reader may also get an impression of the evolution of MR-based schemes during the past fifteen years.

The first MR image fusion approach proposed in the literature is due to Burt [54]. His implementation used a Laplacian pyramid and a sample-based maximum selection rule with $a_S^{(k)}(\cdot) = |y_S^{(k)}(\cdot)|$. Toet [44] presented a similar algorithm but using the ratio-of-low-pass pyramid. His approach is motivated by the fact that the HVS is based on contrast, and therefore, a fusion technique which selects the highest local luminance contrast is likely to provide better details to a human observer. Another variation of this scheme is obtained by replacing the linear filters by morphological ones [21,42].

Burt and Kolczynski [3] proposed to use the gradient pyramid (hence $P = 4$) together with a combination algorithm that is based on an activity and a match measure. In particular, they define the activity of $y_S^{(k)}(\cdot)$ as a local energy measure:

$$a_S^{(k)}(\boldsymbol{n}|p) = \sum_{\Delta\boldsymbol{n}\in\mathscr{W}^{(k)}(p)} \left|y_S^{(k)}(\boldsymbol{n}+\Delta\boldsymbol{n}|p)\right|^2, \tag{20}$$

and the match between $y_A^{(k)}(\cdot)$ and $y_B^{(k)}(\cdot)$ as

$$m_{AB}^{(k)}(\boldsymbol{n}|p) = \frac{2\sum_{\Delta\boldsymbol{n}\in\mathscr{W}^{(k)}(p)} y_A^{(k)}(\boldsymbol{n}+\Delta\boldsymbol{n}|p) y_B^{(k)}(\boldsymbol{n}+\Delta\boldsymbol{n}|p)}{a_A^{(k)}(\boldsymbol{n}|p) + a_B^{(k)}(\boldsymbol{n}|p)}, \tag{21}$$

with $\mathscr{W}^{(k)}(p)$ being either a 1×1, 3×3 or 5×5 window centered at the origin. The combination process is the weighted average $y_F^{(k)}(\cdot) = w_A(d^{(k)}(\cdot))y_A^{(k)}(\cdot) + w_B(d^{(k)}(\cdot))\times y_B^{(k)}(\cdot)$, where the weights are determined by the decision process for each level k, band p and position $\boldsymbol{n}$ as $w_A(d^{(k)}(\cdot)) = 1 - w_B(d^{(k)}(\cdot)) = d^{(k)}(\cdot)$, with

$$d^{(k)}(\cdot) = \begin{cases} 1 & \text{if } m_{AB}^{(k)}(\cdot) \leqslant T \text{ and } a_A^{(k)}(\cdot) > a_B^{(k)}(\cdot), \\ 0 & \text{if } m_{AB}^{(k)}(\cdot) \leqslant T \text{ and } a_A^{(k)}(\cdot) \leqslant a_B^{(k)}(\cdot), \\ \frac{1}{2}+\frac{1}{2}\left(\frac{1-m_{AB}^{(k)}(\cdot)}{1-T}\right) & \text{if } m_{AB}^{(k)}(\cdot) > T \text{ and } a_A^{(k)}(\cdot) > a_B^{(k)}(\cdot), \\ \frac{1}{2}-\frac{1}{2}\left(\frac{1-m_{AB}^{(k)}(\cdot)}{1-T}\right) & \text{if } m_{AB}^{(k)}(\cdot) > T \text{ and } a_A^{(k)}(\cdot) \leqslant a_B^{(k)}(\cdot) \end{cases} \tag{22}$$

for some threshold T. Observe that in case of a poor match (no similarity between the inputs), the source coefficient having the largest activity will yield the composite value, while otherwise, a weighted sum of the sources coefficients will be used. The authors claim that this approach provides a partial solution to the problem of combining components that have opposite contrast, since such components are combined by selection. In addition, the use of area-based (versus sampled-based) operations and the gradient pyramid provide greater stability in noise, compared to the Laplacian pyramid-based fusion.

Ranchin and Wald [55] presented one of the first wavelet-based fusion systems. This approach is also used by Li et al. [35]. Their implementation considers the maximum absolute value within a window as the activity measure associated with the sample centered in the window: $a_S^{(k)}(\boldsymbol{n}|p) = \max_{\Delta\boldsymbol{n}\in\mathscr{W}^{(k)}(p)}\left(\left|y_S^{(k)}(\boldsymbol{n}+\Delta\boldsymbol{n}|p)\right|\right)$. For each position in the transform domain, the maximum selection rule is used to determine which of the inputs is likely to contain the most useful information. This results in a preliminary decision map which indicates, at each position, which source should be used in the combination map. This decision map is then subject to a consistency verification. In particular, Li et al. apply a majority filter in order to remove possible wrong selection decisions caused by impulsive noise. The authors claim that their scheme performs better than the Laplacian pyramid-based fusion due to the compactness, directional selectivity and orthogonality of the wavelet transform.

Wilson et al. [56] used a DWT fusion method and a perceptual-based weighting based on the frequency response of the HVS. Indeed, their activity measure is computed as a weighted sum of the Fourier transform coefficients of the wavelet decomposition, with the weights determined by the contrast sensitivity. [3] They define a perceptual distance between the sources as

$$D_{AB}^{(k)}(\cdot) = \left|\frac{a_A^{(k)}(\cdot) - a_B^{(k)}(\cdot)}{a_A^{(k)}(\cdot) + a_B^{(k)}(\cdot)}\right|,$$

and use this together with the activity to determine the weights of the wavelet coefficients from each source. Observe that this perceptual distance is directly related to the matching measure: the smaller the perceptual distance, the higher the matching measure. The final weighting is given by $w_A\big(d^{(k)}(\cdot)\big) = 1 - w_B\big(d^{(k)}(\cdot)\big) = d^{(k)}(\cdot)$, with

$$d^{(k)}(\cdot) = \begin{cases} 1 & \text{if } D_{AB}^{(k)}(\cdot) > T \text{ and } a_A^{(k)}(\cdot) > a_B^{(k)}(\cdot), \\ 0 & \text{if } D_{AB}^{(k)}(\cdot) > T \text{ and } a_A^{(k)}(\cdot) \leqslant a_B^{(k)}(\cdot), \\ 1-\frac{1}{2}\frac{a_B^{(k)}(\cdot)}{a_A^{(k)}(\cdot)} & \text{if } D_{AB}^{(k)}(\cdot) \leqslant T, \end{cases}$$

for some threshold T. In the experimental results presented by the authors, the fused images obtained with

[3] The contrast sensitivity is defined as the reciprocal of the threshold contrast required for a given spatial frequency to be perceived.

their method are visually better than the ones obtained by fusion techniques based on the gradient pyramid or the ratio-of-low-pass pyramid.

Koren et al. [57] used a steerable wavelet transform for the MR decomposition. They advocate their choice because of the shift-invariance and no-aliasing properties this transform offers. For each frequency band, the activity is a local oriented energy. Only the components corresponding to the frequency band whose activity is the largest are included for reconstruction. Liu et al. [58] also used a steerable pyramid but rather than using it to fuse the source images, they fuse the various bands of this decomposition by means of a Laplacian pyramid.

In [10], Rockinger considered an approach based on a shift-invariant extension of the DWT. The detail coefficients are combined by a maximum selection rule, while the coarse approximation coefficients are merged by averaging. Due to the shift-invariance representation, the proposed method is particularly useful for image sequence fusion, where a composite image sequence has to be built from various input image sequences. The author shows that the shift-invariant fusion method outperforms other MR fusion methods with respect to temporal stability [4] and consistency. [5]

Pu and Ni [59] proposed a contrast-based image fusion method using also the DWT. They measure the activity as the absolute value of what they call directive contrast:

$$a_S^{(k)}(\boldsymbol{n}|p) = \left| \frac{y_S^{(k)}(\boldsymbol{n}|p)}{y_S^{(k)}(\boldsymbol{n}|0)} \right| \quad p = 1, \ldots, 3$$

and use a maximum selection rule as the combination method of the wavelet coefficients. They also proposed an alternative approach where the combination process is performed on the directive contrast itself.

Li and Wang [60] examined the application of discrete multiwavelet transforms to multisensor image fusion. The composite coefficients are obtained through a sample-based maximum selection rule. The authors showed experimental results where their fusion scheme performs better than those based on comparable scalar wavelet transforms.

Another MR technique is proposed by Scheunders [61] where the fusion consists of retaining the modulus maxima [46] of the wavelet coefficients from the different bands and combining them. Noise reduction can be applied during the fusion process by removing noise-related modulus maxima. In the experiments presented, the proposed method outperforms other wavelet-based fusion techniques.

[4] A fused image sequence is temporal stable if the gray level changes in the fused sequence is only caused by the gray level changes in the input sequences.

[5] A fused image sequence is temporal consistent if the gray level changes occurring in the input sequences is present in the fused sequence without any delay or contrast change.

3.3. Experimental results

Here, we give a few examples of fused images. In all cases, we have used the sources shown in Fig. 1, which correspond to visual and IR image modalities. For displaying purposes the gray values of the pixels have been scaled between 0 and 255. Three-levels of decomposition (i.e., $K = 3$) have been used for the MR decomposition of the sources.

Fig. 12 shows some examples of fused images obtained by some of the fusion algorithms which have been discussed before. Fig. 12(a) is the result of a pixel-by-pixel average of the sources. We can observe the lost of contrast compared to the other examples. In Fig. 12(b) we have used Burt's method [54]: a Laplacian pyramid decomposition and the combination algorithm specified in (18) and (19). The same combination algorithm but using a ratio-of-low-pass pyramid for the decomposition (Toet's method [44]) yields the result in Fig. 12(c). Fig. 12(d) illustrates the fusion algorithm proposed by Burt and Kolczynski [3]. Here, the activity measure, match measure and weights are computed as in (20)–(22), with a 3×3 window centered at the origin and a threshold $T = 0.85$.

Fig. 13 shows some examples obtained by combining different alternatives in the different fusion blocks. The purpose is not to outperform the already existing approaches but to give an idea of the flexibility and freedom our framework offers. In Fig. 13(a) a steerable pyramid with $P = 4$ is used for the decomposition of the sources. The combination algorithm is the same as the one proposed by Burt and Kolczynski, with a 3×3 window centered at the origin and a threshold $T = 0.85$. Fig. 13(b) has also been obtained with the same combination algorithm but performing, in addition, consistency check, and using a translation invariant Haar wavelet for the MR representation of the sources. Fig. 13(c) and (d) are examples of fused images where a median pyramid is employed for the MR decomposition. In Fig. 13(d), however, the ratio of the median filtered approximations instead of the standard difference is used. In both cases, the simple combination algorithm specified in (18) and (19) is applied.

4. A region-based MR image fusion algorithm

The algorithms based on MR techniques that we have discussed in the previous sections are mainly pixel-based approaches where each individual coefficient of the MR decomposition (or possibly the coefficients in a small fixed window) is treated more or less independently. However, for most, if not all, image fusion applications, it seems more meaningful to combine objects rather than pixels. As an intermediate step from pixel-based toward object-based fusion schemes, one may

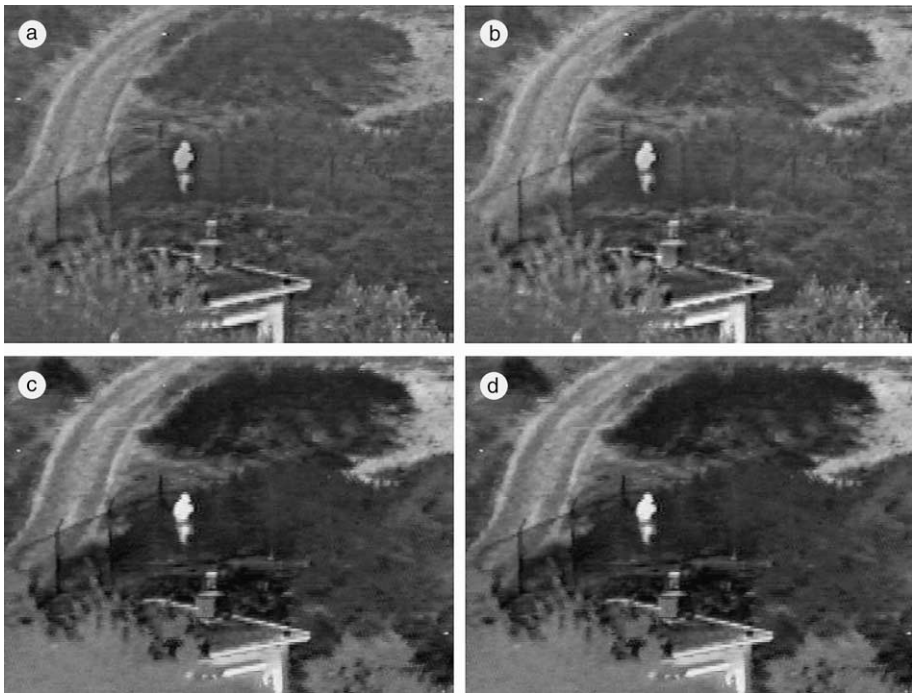

Fig. 12. Examples of fused images by some existing methods: (a) average; (b) Burt's method; (c) Toet's method; (d) Burt and Kolczynski's method.

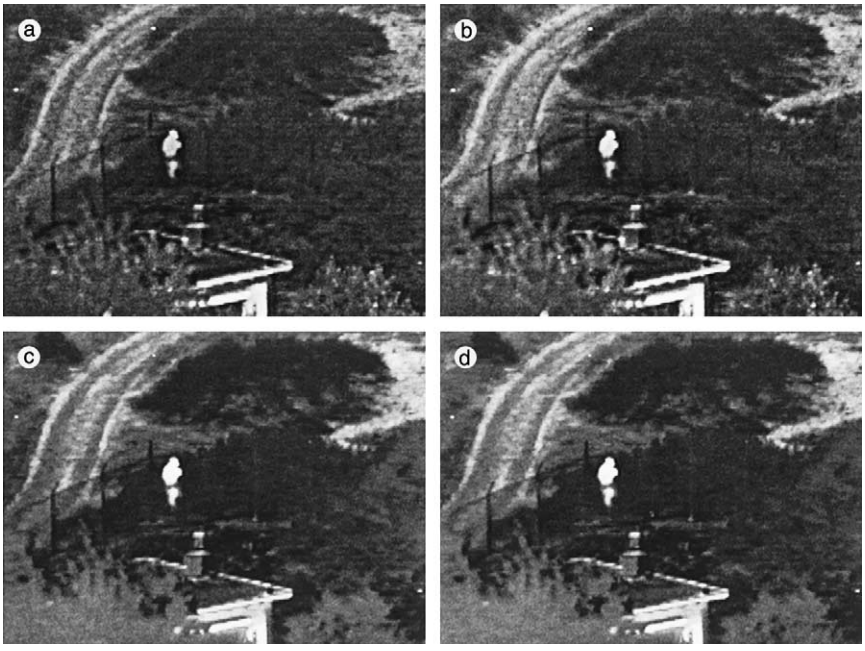

Fig. 13. Examples of fused images by other methods: (a) steerable pyramid with Burt and Kolczynski's combination algorithm; (b) undecimated DWT with Burt and Kolczynski's combination algorithm and consistency check; (c) median pyramid with combination algorithm in (18) and (19); (d) ratio-of-median pyramid with combination algorithm in (18) and (19).

consider region-based approaches. Such approaches have the additional advantage that the fusion process becomes more robust and avoids some of the well-known problems in pixel-level fusion such as blurring effects and high sensitivity to noise and misregistration.

In this section, we introduce a new region-based approach to MR fusion, which combines aspects of feature and pixel-level fusion. The basic idea is to make a segmentation based on all different source images and to use this segmentation to guide the combination process. A major difference with other existing region-based approaches [38,62] is that the segmentation performed is: (i) *multisource*, in the sense that a *single* segmentation is obtained from all the input images, and (ii) *multiresolution*,

in the sense that it is computed in a MR fashion (thus, it is not a segmentation of a sequence of images at different resolutions). For instance, in [38] the regions are obtained by segmenting (independently) each of the approximation images $x_S^{(K)}$ and by exploiting the 'family-relations' we discussed in Section 3.1.5, every detail coefficient $y_S^{(k)}(\cdot)$ is assigned to a region in $x_S^{(K)}$.

4.1. The overall scheme

Our region-based fusion scheme (see Fig. 14) extends the pixel-based fusion approach discussed in Section 3 (see Fig. 11). Indeed, it includes all the blocks described before. The major difference between the two schemes consists hereof that the region-based scheme also contains a segmentation module which uses all sources x_S as input and returns a single MR segmentation $\mathscr{R}$ (i.e., a partition of the underlying image domains into regions) as output. Thus, we use MR decompositions to represent the input images at different scales and, additionally, we introduce a multiresolution/multisource (MR/MS) segmentation to partition the image domain at these scales. The activity and match measures are computed for every region in the decomposed input images. These measures may correspond to low-level as well as intermediate-level structures. Furthermore, the MR segmentation $\mathscr{R}$ allows us to impose data-dependent consistency constraints based on spatial as well as inter- and intra-scale dependencies. All this information, i.e., the measures and the consistency constraints, is integrated to yield a decision map d which governs the combination of the coefficients of the transformed sources. This combination results in a MR decomposition y_F, and by MR synthesis we obtain a fused image x_F.

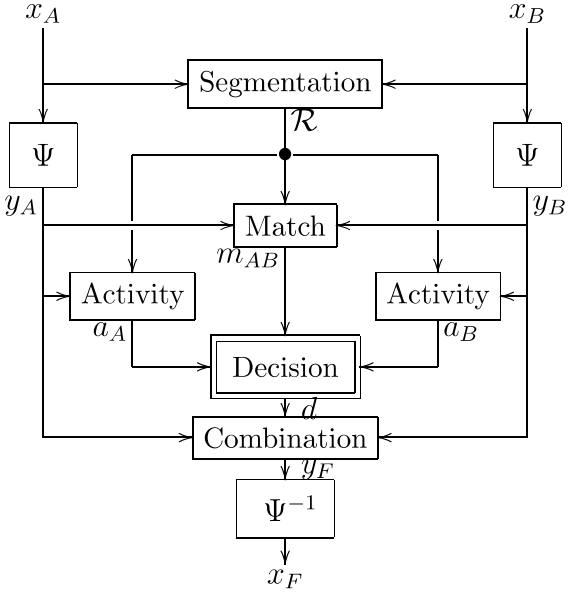

Fig. 14. Generic region-based MR fusion scheme.

The main functional blocks of this fusion strategy are depicted in Fig. 14. Since we already discussed most of them in Section 3, we concentrate on the segmentation module and its interaction with the other modules.

4.1.1. MR/MS segmentation

This block uses the various source images as input and returns a single MR segmentation

$$\mathscr{R} = \{\mathscr{R}^{(1)}, \mathscr{R}^{(2)}, \ldots, \mathscr{R}^{(K)}\}$$

as output. Here $\mathscr{R}^{(k)}$ represents a segmentation at level k, i.e., a partitioning of the domain at level k. Loosely speaking, $\mathscr{R}$ provides a MR representation of the various regions of the underlying scene. This representation will guide the other blocks of the fusion process; hence instead of working at pixel-level, they will take into consideration the regions inferred by the segmentation. From an intuitive point of view, we can regard these regions as the constituent parts of the objects in the overall scene.

In our fusion scheme, the segmentation is merely a preparatory step toward actual fusion. In fact, we are not interested in the segmentation of the images per se, but rather in a coarse partition of the underlying scene. Therefore, the segmentation process does not need to be extremely accurate. For our purposes, we have developed a MR/MS segmentation algorithm based on the *linked pyramid* [63]. We describe our segmentation algorithm in Section 4.2. Obviously, other segmentation methods can be used. We require, however, that the sampling structure in $\mathscr{R}$ is the same as in y_S, so that each partition $\mathscr{R}^{(k)}$ corresponds to a partition of the detail image $y_S^{(k)}(\cdot|p)$.

4.1.2. Combination algorithm

Since the building blocks of the combination algorithm in the region-based approach are essentially the same as in the pixel-based case, the combination algorithms discussed in Section 3 can be easily extended to the region-based approach. For example, we can define the activity of each region $R \in \mathscr{R}^{(k)}$ in $y_S^{(k)}(\cdot|p)$ by

$$a_S^{(k)}(R|p) = \frac{1}{|R|}\sum_{\boldsymbol{n}\in R} a_S^{(k)}(\boldsymbol{n}|p), \tag{23}$$

where $|R|$ is the area of region R. Similarly, we can define the match measure of each region $R \in \mathscr{R}^{(k)}$ in the image bands $y_{\mathrm{A}}^{(k)}(\cdot|p)$ and $y_{\mathrm{B}}^{(k)}(\cdot|p)$ by

$$m_{\mathrm{AB}}^{(k)}(R|p) = \frac{1}{|R|}\sum_{\boldsymbol{n}\in R} m_{\mathrm{AB}}^{(k)}(\boldsymbol{n}|p). \tag{24}$$

Given these measures, the decision map can be constructed in several ways as discussed in Section 3.1.5, with the only difference that $a_S^{(k)}(R|p)$, $m_{\mathrm{AB}}^{(k)}(R|p)$ are used instead of $a_S^{(k)}(\boldsymbol{n}|p)$, $m_{\mathrm{AB}}^{(k)}(\boldsymbol{n}|p)$. For instance, a combina-

tion algorithm based on a maximum selection rule would read

$$y_{\mathrm{F}}^{(k)}(\boldsymbol{n}|p)=\begin{cases} y_{\mathrm{A}}^{(k)}(\boldsymbol{n}|p) & \text{if } a_{\mathrm{A}}^{(k)}(R|p) > a_{\mathrm{B}}^{(k)}(R|p) \\ y_{\mathrm{B}}^{(k)}(\boldsymbol{n}|p) & \text{otherwise} \end{cases}$$

$$\text{for all } \boldsymbol{n} \in R. \tag{25}$$

As in the pixel-based scheme, once the decision map is constructed, the mapping performed by the combination process is determined for all coefficients, and the synthesis process yields the fused image x_{F}. Note that for the particular case in which each region corresponds to a single point $\boldsymbol{n}$, the region-based approach reduces to a pixel-based approach. Thus, the region-based MR fusion scheme extends and generalizes the pixel-based approach and offers a general framework for MR-based image fusion which encompasses most of the existing MR fusion algorithms.

4.2. MR/MS segmentation based on pyramid linking

In this section, we present a MR/MS segmentation algorithm based on pyramid linking. We first review the basics of the conventional pyramid linking segmentation method. Then, we modify and extend this method for the segmentation of several inputs.

4.2.1. The linked pyramid

The linked pyramid structure was first described by Burt et al. [63] (related work can be found in [64–67]). It consists of a MR decomposition of an image with the bottom level containing the full-resolution image and each successive higher level being a filtered/subsampled version derived from the level below it (see Section 2.2). The various levels of the pyramid are 'linked' by means of so-called child–parent relations (see Fig. 15) between their samples (pixels); such child–parent links are established during an iterative processing procedure to be described below.

A conventional linked pyramid is constructed as follows. First, an approximation pyramid is produced by low-pass filtering and sampling. Then, child–parent relations are established by linking each pixel in a level (called *child*) to one of the pixels in the next higher level (called *parent*) which is closest in gray value (or in some other pixel attribute). The attribute values of the parents are then updated using the values of their children. The process of linking and updating is repeated until convergence (which always occurs [63]). Finally (or during the linking process), some pixels are labeled as *roots*. In the simplest case, only the pixels in the top level of the pyramid are roots. Every root and the pixels which are connected to it induce a tree in the pyramid. The leaves of each tree correspond to pixels in the full-resolution image which define a *segment* or *region*. Thus, the linked pyramid provides a framework for an iterative process of image segmentation. For example, in Fig. 15, pixel T is a root which represents in the bottom level a segment composed of pixels a, b, c and d.

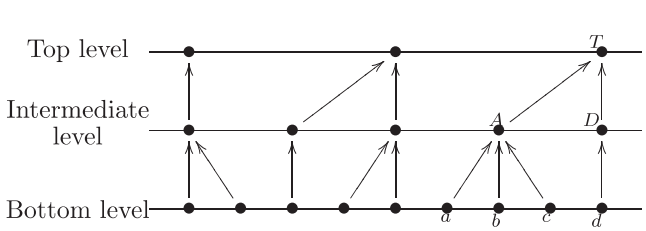

Fig. 15. A diagram illustrating linking relationships: e.g., pixel A is the parent of children a, b and c, and it is also the child of pixel T.

There exist many variations on the scheme: in the way the initial pyramid is built, in the manner pixels are linked to each other, in determining when pixels should be declared as roots, in the size of the neighborhood in which children can look for a parent to link to, in the attribute that is being used (e.g., gray value, edge, local texture), etc.

4.2.2. MR segmentation algorithm using linking

Our basic algorithm follows the classical "50% overlapping 4×4" structure [63]. This means that each parent is derived from the pixels in the 4×4 neighborhood immediately below it, and this neighborhood overlaps 50% of that of its 4 neighbors. Thus, each pixel has 16 candidate children and each child up to 4 candidate parents. The bottom of the pyramid corresponds to level zero and, for simplicity, is assumed to be of size $N \times N$ with N a power of 2. The maximum height of the pyramidal structure is considered to be $K_M = \log_2 N - 1$.

At each level k, the pixels are indexed by the vector $\boldsymbol{n} = (n, m)$, where $n,\ m = 0, \ldots, \frac{N}{2^k} - 1$. We denote by $\mathscr{C}(\boldsymbol{n})$ the set of candidate children of pixel $\boldsymbol{n}$ at level $k > 0$:

$$\mathscr{C}(\boldsymbol{n}) = \{(n', m') | n' \in \{2n-1, 2n, 2n+1, 2n+2, \}, \\ m' \in \{2m-1, 2m, 2m+1, 2m+2\}\}.$$

Similarly, we denote by $\mathscr{P}(\boldsymbol{n})$ the set of candidate parents of pixel $\boldsymbol{n}$ at level $k < K_M$:

$$\mathscr{P}(\boldsymbol{n}) = \Big\{(n', m') | n' \in \big\{\big\lfloor \tfrac{1}{2}(n-1)\big\rfloor, \big\lfloor \tfrac{1}{2}n\big\rfloor, \big\lfloor \tfrac{1}{2}(n+1)\big\rfloor\big\}, \\ m' \in \big\{\big\lfloor \tfrac{1}{2}(m-1)\big\rfloor, \big\lfloor \tfrac{1}{2}m\big\rfloor, \big\lfloor \tfrac{1}{2}(m+1)\big\rfloor\big\}\Big\},$$

where $\lfloor \cdot \rfloor$ denote s the integer part of the enclosed value. The set of pixels to which the pixel $\boldsymbol{n}$ is connected at the bottom level is called *receptive field*.

To each pixel we associate one or more variables representing the attributes on which the segmentation will be based. In this study, we assign to each pixel $\boldsymbol{n}$ at level k its grayscale value $x^{(k)}(\boldsymbol{n})$, and the area $A^{(k)}(\boldsymbol{n})$ of its receptive field.

Consider an input image $x = x^{(0)}$. Our pyramid segmentation algorithm consists of three steps.

1. *Initialization*

   We associate to each pixel $\boldsymbol{n}$ at level zero the gray value $x^{(0)}(\boldsymbol{n})$ of the original image, and to each pixel $\boldsymbol{n}$ at level $k>0$ a gray value $x^{(k)}$ computed from the average of the gray values of its candidate children:
   $$x^{(k)}(\boldsymbol{n}) = \frac{1}{16}\sum_{\boldsymbol{n}'\in\mathscr{C}(\boldsymbol{n})} x^{(k-1)}(\boldsymbol{n}').$$

2. *Linking*

   (a) Pixel linking and root labeling.
   For each child, a suitable parent is sought among the candidate parents: it is linked to its most 'similar' parent or it becomes a root (see below). Here, 'similarity' is based on the difference in grayscale. This difference is computed between the child and each of its four candidate parents. A link is established with the parent that minimizes that distance. If more than one candidate parent minimizes it, we arbitrary pick one of them.
   In our approach, we perform the root labeling within the linking step. That is, when linking to a parent, if the grayscale difference is above some threshold, the link is not established and the pixel is labeled as a root (thus, it is not considered to be a child any more). An advantage of this method is its speed: a single operation will identify all roots. A disadvantage is that it is not clear beforehand how many roots (and, therefore, how many segments) will be found. Defining a good root labeling threshold is not straightforward. When the threshold is too high, few pixels become roots, whereas many pixels are labeled as root if the threshold is too low. For simplicity, we use a threshold $T=0.25\Delta x^{(0)}$ where $\Delta x^{(0)}$ is the length of the dynamic range.
   (b) Updating of area $A^{(k)}$ and gray values $x^{(k)}$.
   The attributes of each parent are recomputed using only the children that are linked to it:
   $$A^{(k+1)}(\boldsymbol{n}) = \sum_{\boldsymbol{n}'\in\mathscr{C}(\boldsymbol{n})} A^{(k)}(\boldsymbol{n}'),$$
   $$x^{(k+1)}(\boldsymbol{n}) = \frac{\sum_{\boldsymbol{n}'\in\mathscr{C}(\boldsymbol{n})} x^{(k)}(\boldsymbol{n}')A^{(k)}(\boldsymbol{n}')}{A^{(k+1)}(\boldsymbol{n})},$$
   where $A^{(0)}(\boldsymbol{n})=1$ for all $\boldsymbol{n}$ at level zero.
   (c) Iteration of (a) and (b) until convergence.

3. *Segmentation*

   The actual segmentation is obtained by using the tree structure of the created links. In each level k, all pixels that are connected to a common root are classified as a single region segment R. In this way, at each level k, we obtain a segmented image $\mathscr{R}^{(k)}$ which contains the different regions R at this level.

4.2.3. MR/MS segmentation algorithm using linking

The segmentation method presented in the last subsection can be extended to the case where we have several input images x_S, $S\in\mathscr{S}$. In this case, the initialization step is performed as before for each image and, in the linking step, the distance between a child $\boldsymbol{n}$ and a candidate parent $\boldsymbol{n}'\in\mathscr{P}(\boldsymbol{n})$ is given by the expression

$$\left(\sum_{S\in\mathscr{S}}\left(x_S^{(k)}(\boldsymbol{n}) - x_S^{(k+1)}(\boldsymbol{n}')\right)^2\right)^{1/2}. \tag{26}$$

As in the scalar case, the candidate $\boldsymbol{n}'$ which minimizes this distance is selected to become the parent unless the distance is above some threshold, in which case $\boldsymbol{n}$ is labeled as a root. Using the new links, the gray values are updated for each $S\in\mathscr{S}$, and the process of linking and updating is iterated until convergence. After the linking step, although the gray values of the various inputs $S\in\mathscr{S}$ will in general differ, the linking structure (child–parent relations) is the same. Thus, we obtain a single linked pyramid structure and we can apply the same segmentation step as before.

We summarize the basic steps of our MR/MS segmentation in the following algorithm.

Algorithm

1. For each input $S\in\mathscr{S}$
   - Construct an approximation pyramid $\{x_S^{(k)}\}$.
2. For each level $k<K_M$
   - While no convergence,
     - For each child $\boldsymbol{n}$ at level k, find the parent $\boldsymbol{n}'\in\mathscr{P}(\boldsymbol{n})$ which minimizes the distance given by (26). If this distance is above some threshold, $\boldsymbol{n}$ is set as a root, otherwise it is linked to $\boldsymbol{n}'$.
     - For each parent $\boldsymbol{n}$ at level $k+1$, update $A^{(k+1)}(\boldsymbol{n})$ and $x_S^{(k+1)}(\boldsymbol{n})$ for all $S\in\mathscr{S}$.
3. For each level k
   - All pixels $\boldsymbol{n}$ at level $k$connected to a common root (or being themselves a root) are classified as a single region segment R (at level k).

The segmentation is based on the approximation pyramids (computed from the grayscale values of the pixels) of the different input sources x_S, which are all treated equally. Obviously this is a very naive approach since different sources may present different amplitude ranges and may not be equally reliable. Thus, prior to segmentation, one might pre-process (e.g., normalization of amplitudes, denoising, etc.) the input images so their attributes become comparable. Alternatively, one can modify the distance measure in (26) and use, for instance,

$$\left(\sum_{S\in\mathscr{S}}\mu_S\left(x_S^{(k)}(\boldsymbol{n}) - x_S^{(k+1)}(\boldsymbol{n}')\right)^2\right)^{1/2},$$

where μ_s is a normalization factor which may depend on several factors such as the dynamic range, noise estimation, entropy, etc. Additionally, the segmentation

algorithm can be improved by the use of connectivity preservation criteria, other root criteria, adaptive windows and probabilistic linking [65–67].

Note that, by construction, the MR segmentation $\mathscr{R}$ obtained with our algorithm has a pyramidal structure where the bottom level is at full resolution (same size as x_S) and each successive coarser level is 1/4 of its predecessor. However, this might not be true for the MR decompositions y_S obtained with the MR analysis block. Note also that the levels from the above MR segmentation $\mathscr{R}$ range from 0 to K_M, whereas the levels from y_S go from 1 to K. In practice, K is smaller than K_M, so we assume henceforth that $K \leqslant K_M$. In addition, we also assume the same lattice structure. Since we require all decompositions to have the same sampling structure, the MR/MS segmentation module should associate to each level k and band p in y_S, the segmentation level k' such that the domain of $y^{(k)}(\cdot|p)$, i.e., $I_y^{(k)}(p)$, has the same dimensions as $\mathscr{R}^{(k')}$. For instance, if y_S corresponds to a Laplacian decomposition, then $k' = k-1$, for $k = 1,\ldots,K$; while if y_S corresponds to a DWT, then $k' = k$ for $k = 1,\ldots,K$ and all $p = 1,\ldots,3$. We assume that these associations are performed inside the MR/MS segmentation module so that we get the output $\mathscr{R} = \{\mathscr{R}^{(1)},\ldots,\mathscr{R}^{(K)}\}$ which has the same sampling structure as $y_S = \{y_S^{(1)},\ldots,y_S^{(K)}\}$.

4.3. Case studies

In this section, we present some experimental results obtained with one of the simplest implementations of the region-based fusion approach.

We consider two input sources x_A and x_B. For their MR decomposition, we use a Laplacian pyramid (thus, we only have a single orientation band, i.e., $P = 1$). We employ the MR/MS segmentation algorithm discussed in Section 4.2. In the combination algorithm, we do not use a matching measure and define the activity of each region $R \in \mathscr{R}^{(k)}$ as in (23), with $a_S^{(k)}(\boldsymbol{n}|p) = |y_S^{(k)}(\boldsymbol{n}|p)|$. The combination process is performed as in (13), with $w_A(\delta) = \delta$ and $w_B(\delta) = 1-\delta$. In the decision process, each component of d is obtained by the following simple decision rules:

- For $p = 0$, $\delta = d^{(K)}(\boldsymbol{n}|0) = \frac{1}{2}$ for all $\boldsymbol{n}$.
- For $p = 1$,
  - for each level k and for each region $R \in \mathscr{R}^{(k)}$:

$$\delta = d^{(k)}(\boldsymbol{n}|1) = \begin{cases} 1 & \text{if } a_A^{(k)}(R|1) > a_B^{(k)}(R|1) \\ 0 & \text{otherwise} \end{cases} \quad \text{for all } \boldsymbol{n} \in R.$$

Note that according to this algorithm, the composite approximation image $y_F^{(K)}(\cdot|0)$ is the pixel-wise average of the approximation images $y_S^{(K)}(\cdot|0)$ and therefore, the region information $\mathscr{R}^{(K)}$ is neglected. The composite detail images $y_F^{(k)}$, however, are constructed by a selective combination as in (25).

We have tested our algorithm on several pairs of images. Three examples are given here to illustrate the fusion process described above. In all cases, $K = 3$ and, when displaying the images, the gray values of the pixels are scaled between 0 and 255 (histogram stretching). The inputs x_A and x_B are shown, respectively, on the left and right top of the corresponding figure. The images are assumed to be registered and no pre-processing is performed. For the decision maps, pixels with $\delta = 0$, 1 are displayed in black and white, respectively. Thus, according to our algorithm, coefficients corresponding to 'white zones' are selected from $y_A^{(k)}$, while coefficients corresponding to 'black zones' are selected from $y_B^{(k)}$.

Fig. 16 shows the fusion of a visible and an IR wavelength images. The first level of the resulting segmentation and decision map are shown in the middle row. The corresponding second levels are displayed on the bottom left. It is interesting to note that, according to $d^{(2)}$, although most of the background is selected from the visual image $y_A^{(2)}$, the region corresponding to the person is selected from the IR image $y_B^{(2)}$. The fused image is depicted at the bottom right of Fig. 16. Note that in the fused image there is less contrast between the person and the background than in the IR image. This is due to the fact that the approximation images at the coarsest level are averaged, i.e., no region information has been used there.

Fig. 17 shows the fusion of images with different focus points. The segmentation and decision for the first level are displayed in the second row of Fig. 17. Note that since the digit '8' is connected to a particular region located within the left clock, the binary decision map $d^{(1)}$ points out, wrongly, to take the '8' from $y_A^{(1)}$ instead from $y_B^{(1)}$. The same happens in level $k = 2$ (not displayed here). For this particular example, we also show the fused image resulting from a pixel-based MR fusion algorithm with the same fusion rules as in the region-based algorithm. We also illustrate how we can improve the region-based fused image by filtering the decision map. Here, we have filtered both decision maps $d^{(1)}$, $d^{(2)}$ with a morphological alternating filter: an opening followed by a closing [43]. The filtered $d^{(1)}$ is shown at the bottom left of Fig. 17. One can see that small white and black regions have been removed and that the boundaries have been smoothed. The fused image obtained with the filtered decision maps is shown at the bottom right of Fig. 17.

Fig. 18 shows the fusion of a magnetic resonance image (MRI) with a computer tomography (CT) image. In this last example, we illustrate the combination of the approximation coefficients using an activity based on a local variance (see below). More precisely, we perform the selective combination (25) for both detail and approximation coefficients but using different activity measures. For the details, $a_S^{(k)}(R|1)$ is defined as before, while for the approximation we consider

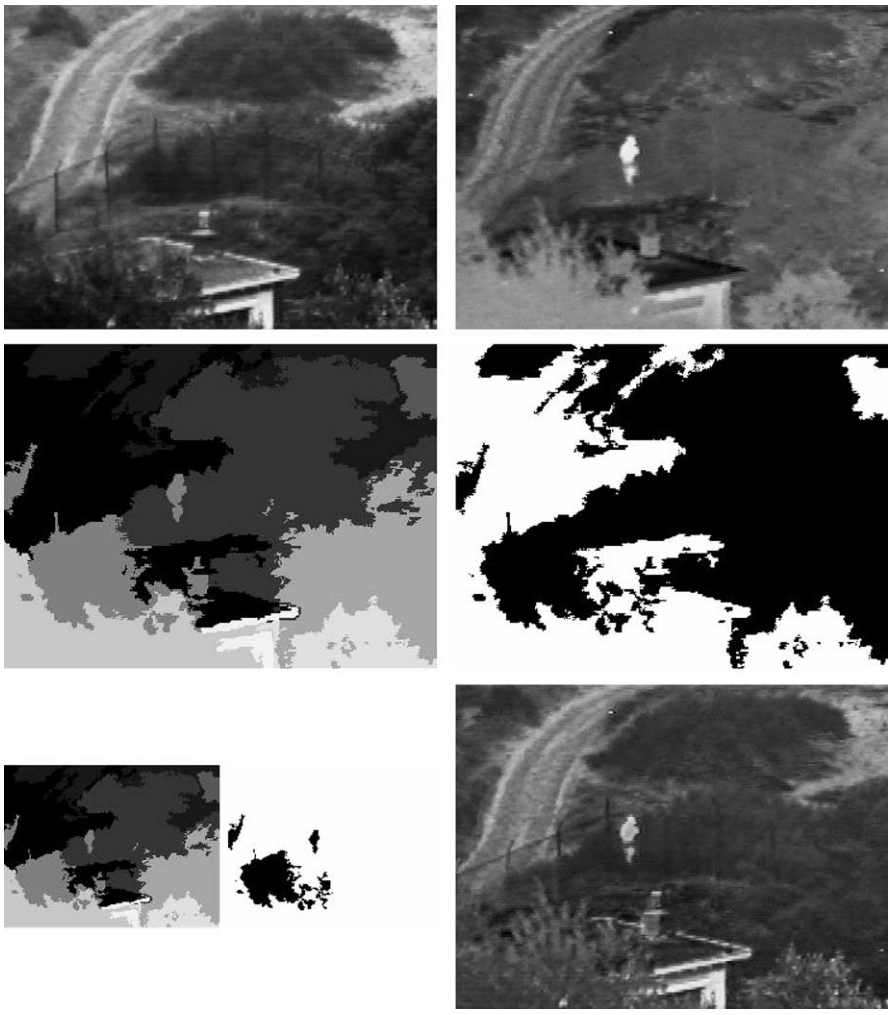

Fig. 16. Example 1. Top: visual (left) and IR (right) test images; middle: first level of segmentation (left) and decision map (right); bottom: second level of segmentation (left) and decision map (central), and fused image (right).

$$a_S^{(K)}(R|0) = \frac{1}{|R|}\sum_{\boldsymbol{n}\in R}(y_S^{(K)}(\boldsymbol{n}|0) - \bar{y}_S^{(K)}(R|0))^2,$$

where $\bar{y}_S^{(K)}(R|0) = \frac{1}{|R|}\sum_{\boldsymbol{n}\in R} y_S^{(K)}(\boldsymbol{n}|0)$. By visual inspection, we can see that the region-based fusion (right image of the third row) preserves the soft tissue of the MRI better than pixel-based fusion (bottom left image). The bottom right image is the image resulting from a fusion algorithm where the region-based approach is only used for the detail images (as we did in the previous experiments).

4.4. Discussion

From the experiments presented, we can see that, despite the crudeness of the current implementation, the visual performance is surprisingly good. This suggests that the region-based approach proposed here can at least be competitive with (but more likely outperform) other MR fusion techniques. The topic of objective performance evaluation is discussed in Section 5, where we give a brief overview of some existing evaluation fusion methods and apply one of them to the fused images obtained in the previous experiments.

The influence on the different parameters is a venue of current research. In particular, our MR/MS segmentation is quite sensitive to the root labeling criteria and the threshold proposed in Section 4.2.2 does not always give a satisfactory segmentation.

Further investigations are necessary for the fine-tuning of parameters as well as for the proper selection of the different ingredients of the scheme. Toward this end, new performance assessment criteria will be developed to evaluate and demonstrate the capacities of the new fusion technique, as well as to compare its performance with others MR fusion schemes. Tests will be carried out based both on objective and subjective criteria.

5. Performance assessment

Performance measures are essential to determine the possible benefits of fusion as well as to compare results

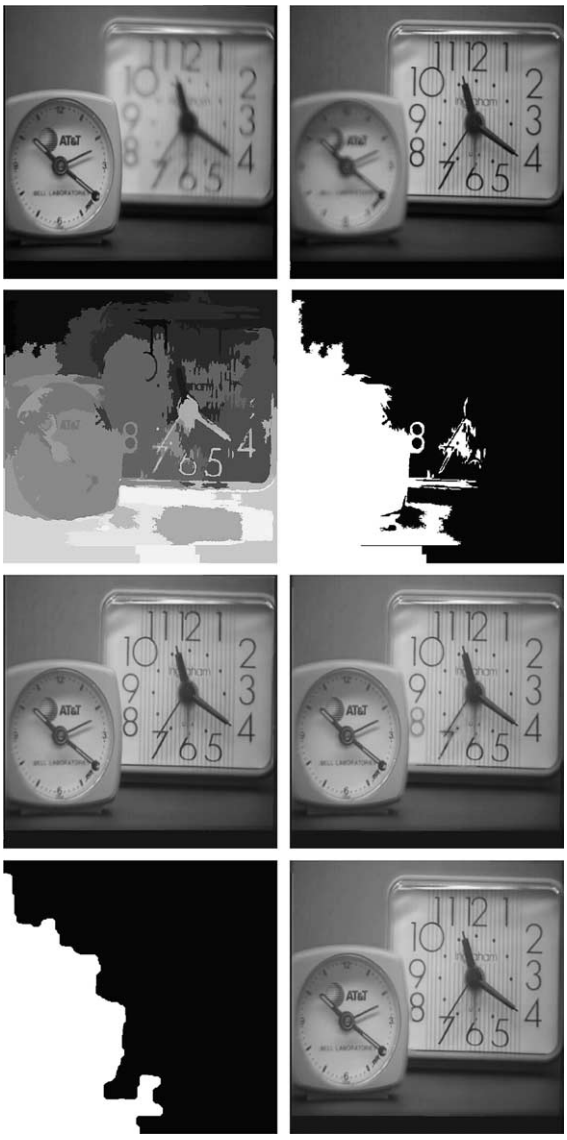

Fig. 17. Example 2. Top: multi-focus test images; second row: first level of segmentation (left) and decision map (right); third row: fused images with pixel-based (left) and region-based (right) approach; bottom: filtered decision map (left) and corresponding fused image.

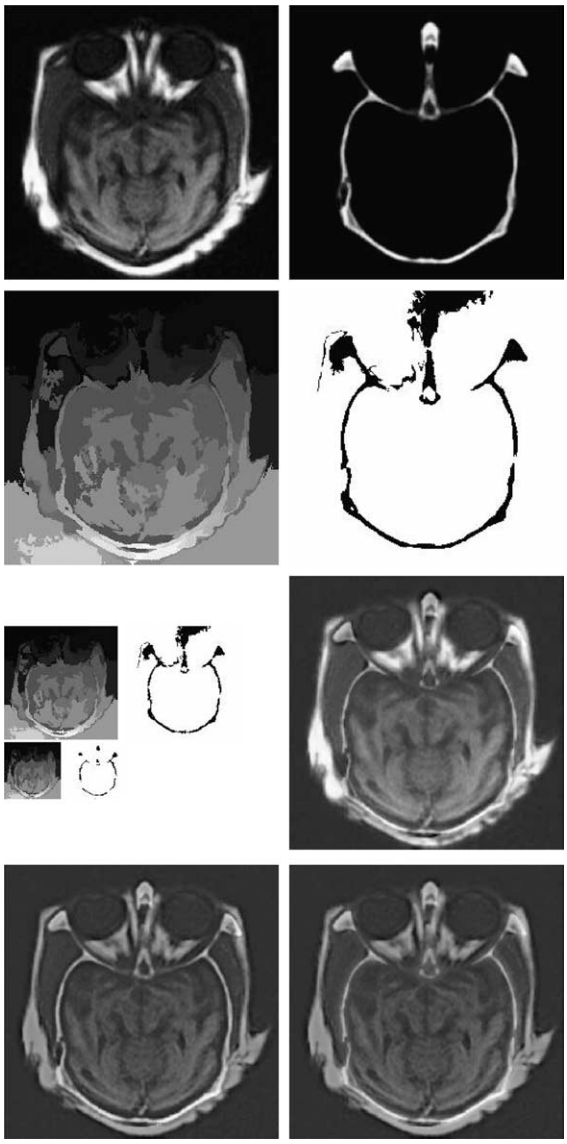

Fig. 18. Example 3. Top: MRI (left) and CT (right) test images; second row: first level of segmentation (left) and decision map (right); third row: second and third level of segmentation (left) and decision map (central), and fused image (right); bottom: pixel-based (left) and region-based (right) fused images with pixel-wise average combination for the approximation images.

obtained with different algorithms. Furthermore, they are necessary in order to obtain an optimal setting of parameters for a specific fusion algorithm.

In many applications, the ultimate user or interpreter of the fused image is a human. Consequently, the human perception of the fused image is of paramount importance and therefore, fusion results are mostly evaluated by subjective criteria [14,15,68]. This involves human observers to judge the quality of the resulting fused images. Since the 'human quality measure' depends highly on psychovisual factors, these subjective tests are difficult to reproduce and verify, as well as time consuming and expensive. Hence, although it cannot be denied that subjective tests are important in characterizing fusion performance, objective performance metrics appear as a valuable complementary method. But how can a subjective impression like image quality be quantified? This problem is usually solved by associating quality with the deviation of the experimental fused image from the 'ideal' fused image. Then, another problem arises, namely, how to define the 'ideal' fused image. A less usual approach is to design performance measures which, without assuming knowledge of a ground-truth, can be used for quality assessment of the fused image. These performances measures quantify the degree to which the fused image is 'related' to the input sources.

The work by Li et al. [35] is an example where out-of-focus image fusion is evaluated by comparison of the fused image with an 'ideal' composite created by a manual 'cut and paste' process. Indeed, various fusion algorithms presented in literature have been evaluated by constructing some kind of ideal fused image and using it as a reference for comparing with the experimental fused results. Mean squared error based metrics are widely used for these comparisons, and despite their well-known limitations, they can be helpful if used carefully. An example of such metrics is the root mean square error (RMSE):

$$\mathrm{RMSE} = \left(\frac{1}{MN} \sum_{n=1}^{N} \sum_{m=1}^{M} (x_{\mathrm{R}}(n,m) - x_{\mathrm{F}}(n,m))^2 \right)^{1/2}, \tag{27}$$

where x_{R} is the ideal reference, x_{F} the obtained fused image, and M, N are the dimensions of the images.

Information-theory related metrics [69] such as *mutual information* have also been proposed for fusion evaluation [37]. Given two images x_{F} and x_{R} we define their mutual information as

$$I(x_{\mathrm{R}}; x_{\mathrm{F}}) = \sum_{u=1}^{L} \sum_{v=1}^{L} h_{\mathrm{R,F}}(u,v) \log_2 \frac{h_{\mathrm{R,F}}(u,v)}{h_{\mathrm{R}}(u) h_{\mathrm{F}}(v)}, \tag{28}$$

where h_{R}, h_{F} are the normalized gray level histograms of x_{R}, x_{F}, respectively, $h_{\mathrm{R,F}}$ is the joint gray level histogram of x_{R} and x_{F}, and L is the number of bins. Let x_{R}, x_{F} correspond to the reference and fused images, respectively; $I(x_{\mathrm{R}}; x_{\mathrm{F}})$ indicates how much information the fused image x_{F} conveys about the reference x_{R}. Thus, the higher the mutual information between x_{F} and x_{R}, the more likely x_{F} resembles the ideal x_{R}.

An example of an objective performance metric which does not assume the knowledge of ground-truth is given by Xydeas and Petrović [70]. In their approach, important visual information is associated with edge information measured for each pixel. Thus, they measure the fusion performance by evaluating the relative amount of edge information that is transferred from the input images to the fused image. Another non-reference objective performance metric is proposed by Qu et al. [71]. They evaluate image fusion by adding the mutual information between the fused image and each of the input images, i.e.,

$$\mathrm{MI} = I(x_{\mathrm{A}}; x_{\mathrm{F}}) + I(x_{\mathrm{B}}; x_{\mathrm{F}}). \tag{29}$$

From the current literature, we can conclude that objective performance assessment is an open problem which has received little attention. Most existing performance assessment methods are low-level, i.e., they act on the pixel-level. High-level methods, i.e., acting on region or even object level are non-existent. More research is required to provide valuable objective evaluation methods for image fusion, in particular, where if concerns region or object-based methods.

Table 1
Performance comparison between region and pixel-based fusion in the various examples of Section 4.3

MI	Pixel-based	Region-based
Example 1 (person)	1.4418	1.5097
Example 2 (clock)	6.8529	7.0070
Example 3 (skull)	2.6833	4.0637

To get an impression of the potential of the region-based versus the pixel-based scheme, we use the quality metric defined in (29) to evaluate the fused images obtained in Section 4.3 where we do not have a ground-truth. The results are shown in Table 1. In Example 2, we use the region-based fused image shown at the bottom of Fig. 17 (which was obtained with a post-processed decision map). In Example 3, we take the region-based fused image depicted in the third row of Fig. 18 (which was computed using the region information also in the approximation). In all cases, one can see that the region-based method gives a higher MI quality than the pixel-based method. A preliminary conclusion therefore is that the region-based scheme outperforms the pixel-based scheme. Moreover, we can infer that the use of region information for the combination of the approximation images as well as for the detail images (as used in Example 3) improves the fused image considerably.

6. Conclusions

In this paper, we have introduced a general framework for MR image fusion. The proposed framework not only encompasses most of the existing MR image fusion schemes, but also allows the construction of new ones, either pixel or region-based approaches.

The region-based MR fusion scheme presented in this paper is an extension of the classical pixel-based schemes. The basic idea is to perform a MR/MS segmentation of the various input images in order to guide the fusion process. For this purpose, we developed a MR/MS segmentation method based on pyramid linking and suggested some combination algorithms which make use of the resulting segmentation. Experimental results have also been shown.

The implementation of our region-based fusion approach is still in a preliminary stage and in the experiments performed we did not attempt to optimize its performance. However, the results obtained so far suggest that our approach may be useful for several image fusion applications. We need to investigate this more thoroughly in the future. In particular, we plan to study the effect of the different parameters and functions in the scheme on the final fusion process. We also intend to

design new combination algorithms and replace the MR/MS segmentation by pyramid linking by some other techniques, such as a hierarchical watershed from mathematical morphology.

A substantial part of our efforts will be devoted to the design of objective measures for fusion performance assessment. We intend to use such measures to evaluate and demonstrate the capacities of our region-based fusion approach, as well as to compare its performance with other MR fusion schemes. We also plan to study how these objective measures can be used to guide the fusion and improve the fusion performance.

Acknowledgements

The author is indebted to Henk Heijmans for his careful review and helpful suggestions. The author would like also to thank A.G. Steenbeek and P.M. de Zeeuw for their help in the software implementation, and to A. Toet for his valuable feedback and for providing some of the test images.

References

[1] C. Pohl, J.L. Genderen, Multisensor image fusion in remote sensing: concepts, methods and applications, International Journal of Remote Sensing 19 (5) (1998) 823–854.
[2] R.C. Luo, M.G. Kay, Multisensor Integration and Fusion for Intelligent Machines and Systems, Ablex Publishing Corporation, 1995.
[3] P.J. Burt, R.J. Kolczynski, Enhanced image capture through fusion, in: Proceedings of the 4th International Conference on Computer Vision, Berlin, Germany, 1993, pp. 173–182.
[4] M.M. Daniel, A.S. Willsky, A multiresolution methodology for signal-level fusion and data assimilation with application to remote sensing, Proceedings of the IEEE 85 (1) (1997) 164–180.
[5] B.V. Dasarathy, Fuzzy evidential reasoning approach to target identity and state fusion in multisensor environments, Optical Engineering 36 (3) (1997) 683–699.
[6] M. Dubuisson, A.K. Jain, Contour extraction of moving objects in complex outdoor scenes, International Journal of Computer Vision 14 (1995) 83–105.
[7] B.V. Dasarathy, Decision Fusion, IEEE Computer Society Press, Los Alamitos, California, 1994.
[8] B. Jeon, D.A. Landgrebe, Decision fusion approach for multitemporal classification, IEEE Transactions on Geoscience and Remote Sensing 37 (3) (1999) 1227–1233.
[9] L.G. Brown, A survey of image registration techniques, ACM Computing Survey 24 (4) (1992) 325–376.
[10] O. Rockinger, Pixel-level fusion of image sequences using wavelet frames, in: Proceedings of the 16th Leeds Applied Shape Research Workshop, Leeds University Press, 1996.
[11] A. Bastiere, Methods for multisensor classification of airborne targets integrating evidence theory, Aerospace Science and Technology 2 (6) (1998) 401–411.
[12] D.D. Sworder, J.E. Boyd, G.A. Clapp, Image fusion for tracking manoeuvring targets, International Journal of Systems Science 28 (1) (1997) 1–14.
[13] D.W. McMichael, Data fusion for vehicle-borne mine detection, in: EUREL Conference on Detection of Abandoned Land Mines, 1996, pp. 167–171.
[14] D. Ryan, R. Tinkler, Night pilotage assessment of image fusion, in: Proceedings of Helmet and Head-Mounted Displays and Symbology Design Requirements, vol. 2465, SPIE, 1995, pp. 50–67.
[15] A. Toet, J.K. Ijspeert, A.M. Waxman, M. Aguilar, Fusion of visible and thermal imagery improves situational awareness, in: Proceedings of SPIE Conference on Enhanced and Synthetic Vision, vol. 3088, 1997, pp. 177–188.
[16] A.R. Mirhosseini, Y. Hong, M.L. Kin, P. Tuan, Human face image recognition: an evidence aggregation approach, Computer Vision and Image Understanding 71 (2) (1998) 213–230.
[17] I. Couloigner, T. Ranchin, V.P. Valtonen, L. Wald, Benefit of the future SPOT-5 and of data fusion to urban roads mapping, International Journal of Remote Sensing 19 (8) (1998) 1519–1532.
[18] D.G. Leckie, Synergism of SAR and visible/infrared data for forest type discrimination, Photogrammetric Engineering and Remote Sensing 56 (1990) 1237–1246.
[19] T. Toutin, SPOT and Landsat stereo fusion for data extraction over mountainous areas, Photogrammetric Engineering and Remote Sensing 64 (2) (1998) 109–113.
[20] D. Hill, P. Edwards, D. Hawkes, Fusing medical images, Image Processing 6 (2) (1994) 22–24.
[21] G.K. Matsopoulos, S. Marshall, Application of morphological pyramids: fusion of MR and CT phantoms, Journal of Visual Communication and Image Representation 6 (2) (1995) 196–207.
[22] S.T.C. Wong, R.C. Knowlton, R.A. Hawkins, K.D. Laxer, Multimodal image fusion for noninvasive epilepsy surgery planning, IEEE Transactions on Computer Graphics and Applications 16 (1) (1996) 30–38.
[23] M.A. Abidi, R.C. Gonzalez (Eds.), Data Fusion in Robotics and Machine Intelligence, Academic Press, San Diego, 1992.
[24] A. Castellanos, J.D. Tardos, Mobile Robot Localization and Map Building: A Multisensor Fusion Approach, Kluwer Academic Publishers, Boston, MA, 2000.
[25] M. Kam, X. Zhu, P. Kalata, Sensor fusion for mobile robot navigation, Proceedings of the IEEE 85 (1997) 108–119.
[26] R.R. Murphy, Sensor and information fusion improved vision-based vehicle guidance, IEEE Intelligent Systems 13 (6) (1999) 49–56.
[27] K.N. Lou, L.G. Lin, An intelligent sensor fusion system for tool monitoring on a machining centre, International Journal of Advanced Manufacturing Technology (13) (1997) 556–565.
[28] J.M. Reed, S. Hutchinson, Image fusion and subpixel parameter estimation for automated optical inspection of electronic components, IEEE Transactions on Industrial Electronics 43 (3) (1996) 346–354.
[29] E. Lallier, Real-time pixel-level image fusion through adaptive weight averaging, Technical Report, Royal Military College of Canada, 1999.
[30] J.A. Richards, Thematic mapping from multitemporal image data using the principal component transformation, Remote Sensing of Environment 16 (1984) 36–46.
[31] R.K. Sharma, M. Pavel, Adaptive and statistical image fusion, Society for Information Display Digest of Technical Papers 27, 1996, pp. 969–972.
[32] T. Fechner, G. Godlewski, Optimal fusion of TV and infrared images using artificial neural networks, in: Proceedings of SPIE, vol. 2492, 1995, pp. 919–925.
[33] A.M. Waxman, A.N. Gove, D.A. Fay, J.P. Racamato, J.E. Carrick, M.C. Seibert, E.D. Savoye, Color night vision: opponent processing in the fusion of visible and IR imagery, Neural Networks 10 (1) (1997) 1–6.

[34] L.J. Chipman, T.M. Orr, Wavelets and image fusion, in: Proceedings of the IEEE International Conference on Image Processing, Washington DC, 1995, pp. 248–251.
[35] H. Li, B.S. Manjunath, S.K. Mitra, Multisensor image fusion using the wavelet transform, Graphical Models and Image Processing 57 (3) (1995) 235–245.
[36] A. Toet, Hierarchical image fusion, Machine Vision Application (1990) 1–11.
[37] Z. Zhang, R. Blum, A categorization of multiscale-decomposition-based image fusion schemes with a performance study for a digital camera application, Proceedings of the IEEE 87 (8) (1999) 1315–1326.
[38] Z. Zhang, R. Blum, A region-based image fusion scheme for concealed weapon detection, in: Proceedings of the 31th Annual Conference on Information Sciences and Systems, 1997, pp. 168–173.
[39] J. Goutsias, H.J.A.M. Heijmans, Nonlinear multiresolution signal decomposition schemes. Part I: Morphological pyramids, IEEE Transactions on Image Processing 9 (11) (2000) 1862–1876.
[40] H.J.A.M. Heijmans, J. Goutsias, Nonlinear multiresolution signal decomposition schemes. Part II: Morphological wavelets, IEEE Transactions on Image Processing 9 (11) (2000) 1897–1913.
[41] P.J. Burt, E.H. Adelson, The Laplacian pyramid as a compact image code, IEEE Transactions on Communications 31 (1983) 532–540.
[42] A. Toet, A morphological pyramidal image decomposition, Pattern Recognition Letters 9 (1989) 255–261.
[43] P. Soille, Morphological Image Analysis, Springer-Verlag, Berlin, 1999.
[44] A. Toet, Image fusion by a ratio of low-pass pyramid, Pattern Recognition 9 (1989) 245–253.
[45] I. Daubechies, Ten Lectures on Wavelets, Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, 1992.
[46] S.G. Mallat, A Wavelet Tour of Signal Processing, Academic Press, San Diego, California, 1998.
[47] J. Kovacević, M. Vetterli, Nonseparable multidimensional perfect reconstruction filter banks and wavelet bases for $\mathbb{R}^n$, IEEE Transactions on Information Theory 38 (1992) 533–555.
[48] V. Strela, N. Heller, G. Strang, The applications of multiwavelets filter banks to signal and image processing, IEEE Transactions on Image Processing 8 (4) (1996) 548–563.
[49] E.P. Simoncelli, W.T. Freeman, The steerable pyramid: a flexible architecture for multi-scale derivative computation, in: Proceedings of the IEEE International Conference on Image Processing, 1995, pp. 444–447.
[50] P.J. Burt, A gradient pyramid basis for pattern selective image fusion, in: Proceedings of the Society for Information Display Conference, 1992.
[51] D. Marr, Vision, W.H. Freeman and Company, 1982.
[52] T. Kohonen, Self-Organizing Maps, Springer-Verlag, 1995.
[53] J.W. Sammon, A nonlinear mapping for data analysis, IEEE Transactions on Computers C 18 (1969) 401–409.
[54] P.J. Burt, The pyramid as a structure for efficient computation, in: A. Rosenfeld (Ed.), Multiresolution Image Processing and Analysis, Springer-Verlag, Berlin, Germany, 1984, pp. 6–35.
[55] T. Ranchin, L. Wald, The wavelet transform for the analysis of remotely sensed images, International Journal of Remote Sensing 14 (1993) 615–619.
[56] T.A. Wilson, S.K. Rogers, L.R. Meyers, Perceptual based hyperspectral image fusion using multiresolution analysis, Optical Engineering 34 (11) (1995) 3154–3164.
[57] I. Koren, A. Laine, F. Taylor, Image fusion using steerable dyadic wavelet transforms, in: Proceedings of the IEEE International Conference on Image Processing, Washington DC, 1995, pp. 232–235.
[58] Z. Liu, K. Tsukada, K. Hanasaki, Y.K. Ho, Y.P. Dai, Image fusion by using steerable pyramid, Pattern Recognition Letters 22 (2001) 929–939.
[59] T. Pu, G. Ni, Contrast-based image fusion using the discrete wavelet transform, Optical Engineering 39 (8) (2000) 2075–2082.
[60] S.T. Li, Y.N. Wang, Multisensor image fusion using discrete multiwavelet transform, in: Proceedings of the 3rd International Conference on Visual Computing, Mexico city, Mexico, 2000.
[61] P. Scheunders, Multiscale edge representation applied to image fusion, in: Proceedings of Wavelet Applications in Signal and Image Processing VIII, SPIE, San Diego, USA, 2000.
[62] B.J. Matuszewski, L.-K. Shark, M.R. Varley, J.P. Smith, Region-based wavelet fusion of ultrasonic, radiographic and shearographyc non-destructive testing images, in: Proceedings of the 15th World Conference on Non-Destructive Testing, Rome, 2000.
[63] P.J. Burt, T.H. Hong, A. Rosenfeld, Segmentation and estimation of image region properties through cooperative hierarchical computation, IEEE Transactions on Systems, Man, and Cybernetics 11 (12) (1981) 802–809.
[64] M. Bister, J. Cornelis, A. Rosenfeld, A critical view of pyramid segmentation algorithms, Pattern Recognition Letters 11 (1990) 605–617.
[65] P.F.M. Nacken, Image segmentation by connectivity preserving relinking in hierarchical graphs structures, Pattern Recognition 9 (6) (1995) 907–920.
[66] M. Spann, C. Horne, Image segmentation using a dynamic thresholding pyramid, Pattern Recognition (22) (1989) 719–732.
[67] K.L. Vincken, A.S.E. Koster, M.A. Vierger, Probabilistic multiscale image segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (2) (1997) 109–120.
[68] A. Toet, E.M. Franken, Perceptual evaluation of different image fusion schemes, Displays 24 (February 2003) 25–37.
[69] T.M. Cover, J.A. Thomas, Elements of Information Theory, John Wiley and Sons, New York, 1991.
[70] C. Xydeas, V. Petrović, Objective pixel-level image fusion performance measure, in: Proceedings of SPIE, vol. 4051, 2000, pp. 88–99.
[71] G.H. Qu, D.L. Zhang, P.F. Yan, Information measure for performance of image fusion, Electronic Letters 38 (7) (2002) 313–315.