
# MULTISCALE IMAGE FUSION

Andreas Ellmauthaler

Tese de Doutorado apresentada ao Programa de Pós-graduação em Engenharia Elétrica, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Engenharia Elétrica.

Orientadores: Eduardo Antônio Barros da Silva
Carla Liberal Pagliari

Rio de Janeiro
Dezembro de 2013

MULTISCALE IMAGE FUSION

Andreas Ellmauthaler

TESE SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA ELÉTRICA.

Examinada por:

---

Prof. Eduardo Antônio Barros da Silva, Ph.D.

---

Prof. Carla Liberal Pagliari, Ph.D.

---

Prof. Hae Yong Kim, D.Sc.

---

Prof. Cláudio Rosito Jung, D.Sc.

---

Prof. Sergio Lima Netto, Ph.D.

RIO DE JANEIRO, RJ – BRASIL
DEZEMBRO DE 2013

Ellmauthaler, Andreas

Multiscale Image Fusion/Andreas Ellmauthaler. – Rio de Janeiro: UFRJ/COPPE, 2013.

XXI, 161 p.: il.; 29, 7cm.

Orientadores: Eduardo Antônio Barros da Silva
Carla Liberal Pagliari

Tese (doutorado) – UFRJ/COPPE/Programa de Engenharia Elétrica, 2013.

Referências Bibliográficas: p. 147 – 160.

1. Image Fusion. 2. Multiscale Transforms. 3. Camera Calibration. I. Silva, Eduardo Antônio Barros da *et al.* II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Elétrica. III. Título.

*Für meine Eltern, Franz und Erika*

# Acknowledgements

Ich möchte mich zuallererst bei meinen Eltern Erika und Franz bedanken. Ohne eure Unterstützung, euren Rückhalt wären die vorliegenden Seiten nicht möglich gewesen. Diese Arbeit ist euch gewidmet. Des Weiteren gilt mein aufrichtiger Dank Eric, Herbert, Jan, Peter, Philipp und Thomas sowie allen Anderen die mich in den letzten Jahren aus der Ferne unterstützt und begleitet haben. Ihr seit der Grund warum ich mich immer nach Österreich zurücksehnen werde.

Gostaria de agradecer aos meus orientadores Prof. Eduardo A. B. da Silva e Prof. Carla L. Pagliari. Sobretudo, pela orientação acadêmica, a amizade e pelo carinho que demostraram por mim ao longo dos últimos anos. Trabalhar com pessoas tão brilhantes e motivadas foi um grande privilégio. A minha vinda ao Brasil não teria sido possível sem o apoio de vocês. Por tudo isso, e muito além, meu muito obrigado.

Aos meus amigos do LPS/SMT pela forma calorosa com a qual me receberam, pela amizade e os momentos de diversão dentro e fora do laboratório. Em especial: Rafael Amado, Prof. Flávio Ávila, Prof. Luiz Wagner P. Biscainho, Fransisco Carlos Jr., Prof. Fabiano Castoldi, Prof. João Dias, Breno Espindola, Ana Fernanda, Prof. Tadeu Ferreira, Jonathan Gois, Camila Gussen, Axel Hollanda, Dr. José Fernando Leite, Alexandre Leizor, Prof. Amaro Lima, Dr. Markus Lima, Amanda Loiola, Luis Lucas, Prof. Wallace Martins, Prof. Gabriel Matos, Michelle Nogueira, Leonardo Nunes, Anderson Oliveira, Prof. Rodrigo Peres, Rodrigo Prates, Prof. Thiago Prego, Felipe Ribeiro, Dr. Michal Simko, Dr. Iker Sobron, Luiz Tavares, Prof. Michel Tcheou, Dr. Rodrigo Torres e Alan Tygel.

Também gostaria de agradecer aos meus amigos queridos da república Álvaro, André, Bruno, Mara, Rodrigo e Túlio tanto pelos momentos de diversão e distração quanto pelas inúmeras discussões divertidas que foram realizadas na nossa sala.

I would also like to thank my Master Thesis supervisor Prof. Kenneth Nilsson who introduced and stirred my interest in image processing. Many concepts used in this thesis are a result of his advice and guidance during my Master in Sweden.

Last but not least I would like to express my deepest appreciation to my girlfriend Daniela. Her unending patience, support and encouragement were essential in the completion of this thesis. Without her help this work would have not been possible. I am indebted to her more than she knows.

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.)

# ALGORITMOS DE FUSÃO DE IMAGENS USANDO DECOMPOSIÇÕES EM MULTIRESOLUÇÃO

Andreas Ellmauthaler

Dezembro/2013

Orientadores: Eduardo Antônio Barros da Silva
Carla Liberal Pagliari

Programa: Engenharia Elétrica

Nas ultimas duas décadas, a comunidade de processamento de imagens acompanhou o surgimento de um novo campo de pesquisa denominado fusão de imagens. O termo se refere ao processo de integração de informações complementares e redundantes de diversas imagens com o objetivo de produzir uma imagem final capaz de descrever uma cena melhor do que as imagens individuais.

Este trabalho tem o intuito de apresentar algoritmos avançados de fusão de imagens usando decomposições em multiresolução. A partir da análise de desempenho de tais decomposições, dois novos métodos de fusão são apresentados. No primeiro método, mostramos que os resultados podem ser significativamente aperfeiçoados após a aplicação de uma nova técnica de fusão, que divide o processo de decomposição de imagens em duas operações sucessivas de filtragem. Além disso, introduzimos uma nova classe de bancos de filtros, que exibem propriedades úteis, como uma robustez elevada contra artefatos de "ringing", por exemplo.

Para guiar a fusão de imagens infravermelhas e visíveis, o primeiro sistema de fusão opera somente ao nível de pixel enquanto o segundo utiliza informação ao nível de regiões. Mais especificamente, através de um novo algoritmo de segmentação, incluímos informações sobre a presença de alvos nas imagens infravermelhas. Diante disso, asseguramos que a informação mais relevante destas imagens é preservada na imagem fundida.

Devido à frequente falta de imagens fontes, a última parte deste trabalho propôs uma nova técnica de registro de imagens infravermelhas e visíveis. Utilizando esta técnica, foi criado um banco de imagens e vídeos que poderá ser utilizado para testar futuras técnicas de fusão de imagens.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

MULTISCALE IMAGE FUSION

Andreas Ellmauthaler

December/2013

Advisors: Eduardo Antônio Barros da Silva
Carla Liberal Pagliari

Department: Electrical Engineering

Over the past two decades, image fusion has emerged as a new and exciting field of research within the image processing community. Generally speaking, image fusion refers to the process of integrating complementary and redundant information from multiple images into one composite image that describes a given scene better than any of the individual source images.

In this work, we aim at providing new and improved image fusion algorithms by means of multiscale transforms. Based on the performance analysis of a variety of multiscale transforms, two novel fusion methods are proposed. In the first approach we show that results can be significantly improved using a novel fusion strategy which splits the image decomposition process into two successive filter operations using spectral factorization of the analysis filters. Moreover, we will introduce a new class of filter banks which exhibits useful properties such as being more robust to ringing artifacts introduced during the fusion process.

Whereas the first fusion system operates solely on the pixel-level, the second one employs region-level information to guide the fusion of infrared-visible image pairs. More specifically, by means of a novel infrared segmentation algorithm, we include information about the presence of targets within the infrared image to the fusion process.

Motivated by the frequent lack of source images, the final part of this work introduces a novel spatiotemporal registration technique for infrared-visible images. By means of the proposed methodology an image and video database was created which can be used by the research community to test and assess novel fusion schemes.

# Contents

# List of Figures

# List of Tables

# List of Abbreviations

| | |
|---|---|
| 1-D | One Dimensional, p. 33 |
| 2-D | Two Dimensional, p. 35 |
| 3-D | Three Dimensional, p. 110 |
| CT | Computed Tomography, p. 5 |
| CVT | Curvelet Transform, p. 9 |
| ConT | Contourlet Transform, p. 9 |
| DFB | Directional Filter Bank, p. 39 |
| DLT | Direct Linear Transformation, p. 115 |
| DTCWT | Dual-Tree Complex Wavelet Transform, p. 6 |
| DWT | Discrete Wavelet Transform, p. 9 |
| EMD | Empirical Mode Decomposition, p. 9 |
| FFT | Fast Fourier Transform, p. 38 |
| ICA | Independent Component Analysis, p. 9 |
| IR | Infrared, p. 3 |
| KLD | Kullback-Leibler Distance, p. 105 |
| KLT | Karhunen-Loève Transform, p. 7 |
| LP | Laplacian Pyramid, p. 38 |
| MI | Mutual Information, p. 47 |
| MRE | Mean Reprojection Error, p. 112 |
| MRI | Magnetic Resonance Imaging, p. 5 |

| | |
|---|---|
| NSCT | Nonsubsampled Contourlet Transform, p. 6 |
| PCA | Principal Component Analysis, p. 5 |
| PDF | Probability Density Function, p. 105 |
| RANSAC | Random Sample Consensus, p. 12 |
| S$\alpha$S | Symmetric Alpha-stable, p. 104 |
| SIFT | Scale-Invariant Feature Transform, p. 11 |
| SSIM | Structural Similarity, p. 49 |
| UWT-SF | UWT with Spectral Factorization, p. 106 |
| UWT | Undecimated Wavelet Transform, p. 9 |
| q-shift | Quarter Sample Shift, p. 45 |

# Chapter 1

# Introduction

Within the last decades substantial progress was achieved in the imagery sensor field. Improved robustness and increased resolution of modern imaging sensors and, more importantly, cheap fabrication costs have made the use of multiple sensors common in a wide range of imaging applications. This development led to the availability of a vast amount of data, depicting the same scene coming from multiple sensors. However, the subsequent processing of the gathered sensor information can be cumbersome since an increase in the number of sensors automatically leads to an increase in the raw amount of sensor data which needs to be stored and processed. This means that longer execution times have to be accepted or the number of processing units and storage devices has to be increased, leading to solutions which may be quite expensive. In addition, when imaging systems are operated by humans, presenting various images may be an overwhelming task for a single observer and may lead to a significant performance drop [4].

One solution for these problems is to replace the entire set of sensor information by a single composite representation which incorporates all relevant sensor data. In image-based applications this plethora of techniques became generally known as image fusion and is nowadays a promising research area.

Image fusion can be summarized as the process of integrating complementary and redundant information from multiple images into one composite image that contains a 'better' description of the underlying scene than any of the individual source images could provide. Hence, the fused image should be more useful for visual inspection or further machine processing [5]. Nevertheless, fusing images is often not a trivial process, since: a) the source images may come from different types of sensors (e.g. with different dynamic range and resolution); b) they tend to exhibit complementary information (e.g. features which appear in some source images but not in all) or c) they may show common information but with reversed contrast, which significantly complicates the fusion process. Furthermore, a fusion approach which is independent of a priori information about the inputs and produces

a composite image that appears 'natural' to a human interpreter is highly desirable. In general, the following requirements can be imposed on the fusion algorithm [6, 7]:

- it should preserve all relevant information contained in the input images;
- it should not introduce any artifacts or inconsistencies which can distract or mislead a human observer or any subsequent image processing task;
- it should be reliable, robust and tolerant of imperfections such as noise and misregistrations.

Image fusion may be applied to images coming from different sensors (multisensor fusion), taken at different times (multitemporal fusion), obtained using various focal lengths (multifocus fusion), taken from different viewpoints (multiview fusion) or captured under different exposure settings (multiexposure fusion).

## 1.1 Categorization of image fusion

The process of image fusion can be performed at three different levels of information representation, namely pixel-, region- or decision-level [5]. In the following we briefly introduce each one of them.

**Pixel-level image fusion**

Image fusion at pixel-level represents the combination of information at the lowest level of information representation, since each pixel in the fused image is determined by a set of pixels in the source images. Usually, this set consists of a single pixel or comprises of all pixels within a small window, typically of size $3 \times 3$ or $5 \times 5$.

The advantage of pixel-level fusion, apart from its easy and time-efficient implementation, is that the resulting image contains the original information from the sources [7]. However, since pixel-level fusion methods are very sensitive to misregistration, co-registered images at subpixel accuracy are required. Today, most image fusion applications employ pixel-level fusion methods.

**Region-level image fusion**

Region-level fusion approaches typically start by extracting all salient features from the various input images. This is done by applying an appropriate segmentation algorithm which identifies all salient features within the input images with respect to certain properties such as size, shape, contrast, texture or gray-level. Based on this segmentation, a region map is created which links each pixel to a corresponding feature. Consequently, the fusion process is performed on the extracted regions

(as opposed to pixel-level image fusion where the fusion result is determined by an arbitrary set of pixels).

Region-level image fusion usually yields advantages compared to pixel-based techniques since some drawbacks, such as blurring effects, high sensitivity to noise and misregistration can be avoided [7]. However, the final fusion performance of region-level image fusion methods highly depends on the quality of the segmentation process. In other words, segmentation errors such as under- or over-segmentation may lead to the absence or degradation of certain features in the fused image [8].

**Decision-level image fusion**

Fusion at decision-level allows the information from multiple sensors to be effectively combined at the highest level of abstraction. In this context, first a decision map is built for each source image by performing a decision (labeling) procedure on all input pixels. Finally, a fused decision map is constructed based on the individual decision maps. For this purpose decision rules are used which reinforce common interpretation and are able to resolve differences between the individual decision maps [7, 9]. ■

The choice of the appropriate level depends on many different criteria such as the underlying application, the characteristics of the physical sources as well as on other factors such as execution time and the available tools. However, there exists a strong inter-linkage between the different levels of image fusion. Many fusion rules which are used to determine the individual pixels in the composite image at pixel-level can, for instance, also be used at region-level to fuse the extracted features. Furthermore, decision-level fusion often resorts to the segmentation map created at region-level to aid with decision-making. In this work we are mainly concerned with the fusion of images at pixel-level. However, in Chapter 5 we introduce a fusion framework which uses concepts of both pixel- and region-level fusion to merge visible and infrared (IR) images.

## 1.2 Application fields

Image fusion has attracted a great deal of attention in a wide variety of different application areas in the last decades. Generally speaking, all imaging applications that require the analysis of more than one image can benefit from image fusion. In what follows we try to classify all these applications into the four main categories: military, remote sensing, medical science and industrial applications.

(a) (b) (c)

**Figure 1.1:** *Military fusion example. (a) Visible image. (b) Infrared image. (c) Resulting image using the fusion framework of Chapter 5 with target detection and target highlighting. Source images kindly provided by Mr. David Dwyer from OCTEC Limited.*

### Military

Historically, military appeared as one of the first application areas for image fusion. It covers applications such as concealed weapon detection [10–13], identification, detection and tracking of targets [14, 15], mine detection [16] and tactical situation assessment [17].

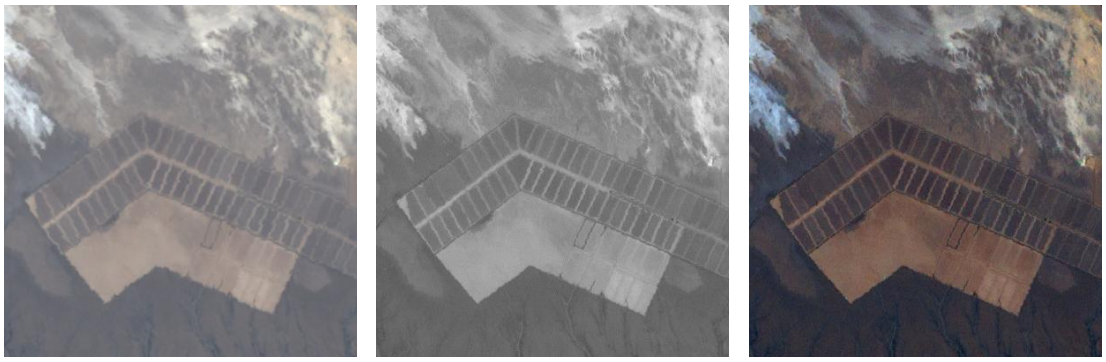
Fig. 1.1 illustrates how the fusion of an IR and visible image pair can be utilized to improve the situation awareness at a location with heavy smoke concentration. It can be noticed that the visible image in Fig. 1.1(a) exhibits a high degree of textural information but is not able to penetrate the smoke. On the other hand, the IR image in Fig. 1.1(b) is able to "see through" the smoke but lacks most of the details depicted in the visible image. The fused image, however, is able to provide the most salient[1] information from both source images.

### Remote sensing

Remote sensing is defined as the measurement of object properties on the earth's surface using data acquired from aircrafts and satellites by means of optical sensors. These systems, particularly those deployed on satellites, provide a repetitive and consistent view of the earth providing valuable information about short- and long-term changes and the impact of human activities [18].

In most remote sensing applications, due to physical constraints, a trade-off between spectral and spatial resolution has to be accepted. In other words, some satellite sensors supply the spectral bands needed to distinguish some features spectrally but not spatially (multispectral image), whereas other sensors include the spatial information needed to distinguish features spatially but not spectrally (panchromatic image). In the context of image fusion we are interested in means to merge images from various sensors into a single image which provide, both, a high spatial and spectral resolution.

[1] We define saliency in this context as the "most relevant information with respect to the underlying application"



(a) (b) (c)

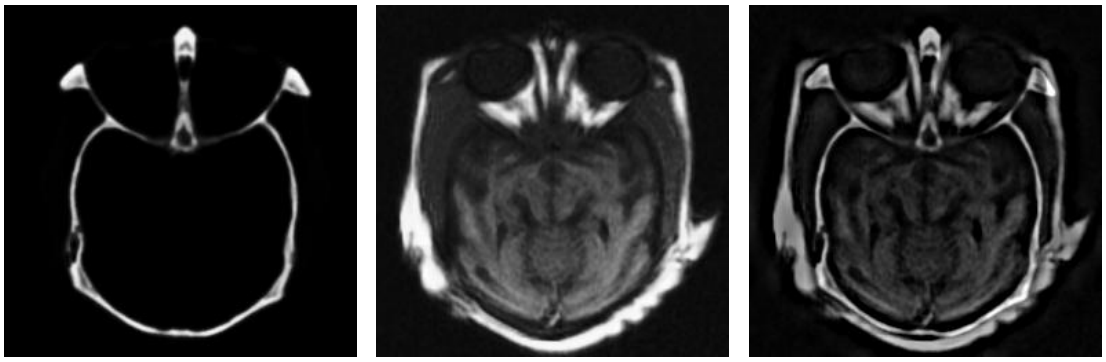
**Figure 1.2:** *Remote sensing fusion example. (a) Multispectral image (pseudo-color). (b) Panchromatic image. (c) Resulting image using a PCA-based fusion strategy (pseudo-color). Source images are taken from* `http://www.AmericaView.org`.

Many image fusion methods have been proposed for this purpose, among them the intensity-hue-saturation transform [19], the Brovey transform [20, 21] and the Principal Component Analysis (PCA) [20, 21] as well as approaches based on multiscale transforms [20–25]. Fig. 1.2 exemplifies the fusion of a multispectral image (Fig. 1.2(a)), consisting of four spectral bands, with the panchromatic image of Fig. 1.2(b), using the PCA method as explained in [20]. It can be observed that the composite image in Fig. 1.2(c) provides more spatial information than Fig. 1.2(a) without losing spectral information. Note that for displaying purposes, Fig. 1.2(a) and Fig. 1.2(c) show only the first three spectral bands in the RGB color space, resulting in the depicted pseudo-color images.

### Medical science

Within the medical community, image fusion has gained an increasing amount of attention in the last decade. Its main application areas can be found in clinical applications such as medical diagnostics, treatment planning and during curative phases such as guided/assisted surgical procedures. The set of input data covers imaging sensors such as Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Positron-Emission-Tomography, Single Photon Emission Computed Tomography, Ultra-Sound and many variants thereof [26].

The fusion of a sample CT and MRI image pair is shown in Fig. 1.3. Here, the information provided by the CT image in Fig. 1.3(a) and the MRI image of Fig. 1.3(b) is complementary. It is well established that soft tissues are better visualized in MRI images than in CT images. Thus, MRI images are commonly used to diagnose pathological soft tissues such as brain tumors. However, the spatial accu-



(a) (b) (c)

**Figure 1.3:** *Medical fusion example. (a) CT image. (b) MRI image. (c) Resulting image using a fusion strategy based on the Dual-Tree Complex Wavelet Transform (DTCWT). Source images kindly provided by Dr. Oliver Rockinger.*

racy of the MRI image for stereotactic[2] localization (e.g. localization of the tissue bone in stereotactic surgery) is very poor due to magnetic susceptibility effects and may result in geometric shifts and distortion effects of up to 4 mm [27]. On the other hand, CT imagery does not suffer from this shortcoming. The fusion of CT and MRI images, as illustrated in Fig. 1.3(c), can therefore be used to remove the geometric distortions inherent in MRI imagery and improve the results in stereotactic radiotherapy.

### Industrial engineering

Image fusion is used in a wide variety of industrial and civil applications. In robotics, multisensor information is used to estimate the position and orientation [28, 29] as well as to navigate a robot in order to avoid collisions and stay on a preset path [28]. Moreover, image fusion is applied in computerized quality management for defect inspection of products [30, 31].

Fig. 1.4 shows an example how image fusion can be used to extend the depth-of-focus of existing image capturing systems. Due to the limited depth-of-focus of individual optical lenses (see Figs. 1.4(a) and 1.4(b)), it is often impossible to get a single image with all objects in focus. One way to overcome this problem is to collect several images from the same scene but with different focus points and combine them into a single composite image which contains the focused regions of all input images.

Another application of image fusion in the industrial context is the combination of multiexposure images [32–34]. A natural scene often has a high dynamic range that exceeds the capture range of common digital cameras. Therefore, a single

[2]stereotactic methods refer to surgical techniques for precisely directing the tip of a delicate instrument (e.g. needle) or beam of radiation in three planes using coordinates provided by medical imaging in order to reach a specific locus in the body

(a) (b) (c)

**Figure 1.4:** *Industrial fusion example. (a) image with focus on the front. (b) image with focus on the back. (c) Resulting image using a fusion strategy based on the Nonsubsampled Contourlet Transform (NSCT). Source images kindly provided by Dr. Oliver Rockinger.*

captured image is usually insufficient to reveal all the details due to under- or over-exposed regions. To solve this problem, images of the same scene can first be captured under different exposure settings and then be combined into a single image using image fusion techniques. ■

The presented list is by no means exhaustive and should merely provide an insight into the most important developments in the field of image fusion. Furthermore, we would like to point out that the image fusion research community is still very active, thus, new application fields are still explored.

In this work we will be restricted to the fusion of multisensor images, exhibiting a high degree of diverging information. Hence, our main focus is placed on the fusion of IR-visible and CT-MRI image pairs as found in military and medical applications.

## 1.3 Fusion techniques

A variety of different image fusion approaches have been developed since the late 80s. In the following we give an excerpt on the most common approaches found in the literature. We divided them into the two groups, transform domain techniques and spatial domain techniques.

### 1.3.1 Transform domain techniques

Transform domain techniques map (transform) each source image into the transform domain (e.g. wavelet or Karhunen-Loève transform (KLT) domain), where the actual fusion process takes place. The final fused image is obtained by taking the inverse transform of the composite representation. The main motivation behind moving

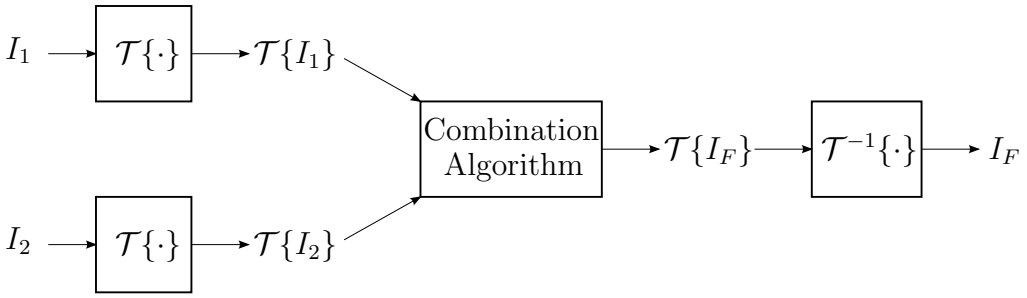



**Figure 1.5:** *Generic fusion scheme in the transform domain according to eq. (1.1).*

to a transform domain is to work within a framework where the image's salient features are more clearly depicted than in the spatial domain. If we let $\mathcal{T}\{\cdot\}$ and $\mathcal{T}^{-1}\{\cdot\}$ represent the forward and inverse transform, respectively, and assume that $g(\cdot)$ represents a function which governs the combination of the (transformed) input images $I_k$, $k = 1, \ldots, K$, commonly known as the "fusion rule", transform domain techniques can be defined as [35]

$$I_F[m,n] = \mathcal{T}^{-1}\{g(\mathcal{T}\{I_1[m,n]\}, \ldots, \mathcal{T}\{I_K[m,n]\})\}, \tag{1.1}$$

where $m, n$ represents the spatial location in the input images and the fused image $I_F$. Fig. 1.5 illustrates this process for two input images. For the purpose of image fusion, transform domain techniques can be roughly categorized into color space transforms, KLT-like transforms and multiscale transforms.

**Color space transforms**

Image fusion by color space transforms takes advantage of the possibility of representing data in different color channels. In the simplest case the individual source images are mapped to a particular color channel (e.g. to the red, green or blue color channel in the RGB space), resulting in a pseudo-color image. These techniques belong to the most frequently used image fusion methods in remote sensing applications. Commonly utilized transforms are the IHS and the Brovey transform [19–21]. In general, color spaces are only defined for three different bands, restricting its use to applications with at most three input images. Often, to escape from this restriction, PCA-based fusion methods are used which allow for an arbitrary number of bands.

**KLT-like transforms**

In order to uncover the underlying structure of an image, it is common practice in image processing applications to represent a source image as the synthesis of several basis images. For this purpose transforms such as the Fourier Transform, the Cosine Transform or the Wavelet Transform have been developed to decompose

an input image using a fixed set of basis images. KLT-like transforms, on the other hand, permit the decomposition of an input image using a basis image set which is especially tailored to the input data. Furthermore, the basis image set can be chosen to be optimal in some statistical sense. For example, the resulting basis images may be desired to be uncorrelated (PCA) or statistically independent, as in case of the Independent Component Analysis (ICA).

In [35] the authors use ICA bases to fuse multifocus and multisensor imagery. In their approach, in order to find proper basis images, the ICA is performed on a set of images with similar content than the ones that will be used for image fusion. After decomposing the actual input images using the previously calculated ICA basis, the fusion is performed in the transform domain in a similar manner as depicted in eq. (1.1). A conceptually similar ICA-based fusion approach is described in [36] for the fusion of IR/visible image pairs.

Fig. 1.2 shows the result for the fusion of multispectral and panchromatic imagery in the context of remote sensing using a PCA-based fusion approach.

**Multiscale transforms**

Among the transform domain techniques, the most frequently used methods are based on multiscale transforms where fusion is performed on a number of different scales and orientations, independently. The multiscale transforms usually employed are Pyramid Transforms [17, 32, 37, 38], the Discrete Wavelet Transform (DWT) [5, 7, 10, 39–43], the Undecimated Wavelet Transform (UWT) [5, 6, 22, 23, 44, 45], the Dual-Tree Complex Wavelet Transform (DTCWT) [8, 46, 47], the Curvelet Transform (CVT) [24, 25, 48], the Contourlet Transform (ConT) [49] and the Non-subsampled Contourlet Transform (NSCT) [50–52]. Due to their importance, the subsequent chapters provide a more detailed view on the use of multiscale transforms in image fusion. ■

Another transform which gained increasing popularity within the image fusion community, and somehow does not fit into any of our categories, is the Empirical Mode Decomposition (EMD). The EMD decomposes a given data set into a number of basis functions, called intrinsic mode functions, which are derived directly from the data. Since the decomposition is based on the local spatial-domain characteristics of the source data, no harmonic analysis is necessary and, thus, it is also applicable to nonlinear and non-stationary processes [53]. Examples for the use of the EMD in image fusion can be found in [54], [55] and [56].

### 1.3.2 Spatial domain techniques

As for spatial domain techniques, the fusion is performed by combining all input images in a linear or non-linear fashion. If we let $g(\cdot)$ represent the chosen fusion rule as in eq. (1.1), spatial domain techniques can be defined as [35]

$$I_F[m,n] = g(I_1[m,n], \ldots, I_K[m,n]). \tag{1.2}$$

In general, spatial domain techniques can be divided into weighted averaging-based, optimization-based and artificial neural network-based approaches.

**Weighted averaging**

A straightforward approach to image fusion is to take the pixel-by-pixel average of the source images. This method, however, leads to undesirable side effects such as reduced contrast as can be observed in Fig. 2.1. A more sophisticated approach to image fusion is to compute each pixel in the composite image as a weighted superposition of all source images. The optimal weighting coefficients can be determined e.g. by performing the PCA of the covariance matrix of the source images [57]. In this approach, the weights for each input image are obtained from the eigenvector corresponding to the largest eigenvalue. Another example is given in [58] where the authors used adaptive weighted averaging for the fusion of IR and visible images.

**Probabilistic fusion**

Probabilistic fusion approaches are based on an image formation model which considers the various input images as being noisy, linearly-transformed versions of an underlying, true scene. The most common image formation model which relates the true scene $I_0$ to the measured, source images $I_k$, $k = 1, \ldots, K$, is given by [59]

$$I_k[m,n] = \beta_k[m,n] I_0[m,n] + \varepsilon_k[m,n], \tag{1.3}$$

where $\beta_k[m,n]$ and $\varepsilon_k[m,n]$ are the gain and sensor noise, respectively, of the $k^{\text{th}}$ sensor at pixel location $m,n$. Thus, the fusion goal is to estimate $I_0$ from $I_k$, $k = 1, \ldots, K$, and can be seen as an inverse problem with eq. (1.3) used as the forward model.

In order to estimate the parameters several approaches can be found in the literature. Sharma et al. [59] used a Bayesian approach whereas in [12] the authors proposed an Expectation-Maximization-based algorithm to estimate the fused image. More recently, Kumar and Dass [60] employed a total variation-based algorithm in conjunction with PCA to estimate the model parameters. Another interesting approach is presented in [61], where Bootstrap sampling in combination with the

Non-parametric Expectation-Maximization algorithm is used.

**Artificial neural networks**

Inspired by the fusion of different sensor signals in biological systems, several researchers have used neural networks to perform image fusion tasks. One of the most famous examples for sensor fusion in a living organism is the visual system of rattlesnakes [62]. These snakes possess an organ which is sensitive to thermal radiation. The IR signals provided by these organs are combined with the nerve signals obtained from the visual sensors, yielding a unique wide-spectrum image of the rattlesnakes environment.

Most fusion techniques employing artificial neural networks concentrate on multifocus image fusion [63–65]. Nevertheless, alternatives exist. In [66], for example, the authors propose the use of a modified pulse-coupled neural network for the fusion of medical image pairs. A further pulse-coupled neural network-based fusion framework for the detection of objects of interest in medical and radar images is introduced in [67]. ■

Henceforth, we confine our discussion to image fusion approaches based on multiscale transforms.

## 1.4 Image Registration

As observed before, an important pre-processing step in image fusion is image registration. In a nutshell, image registration is the process of overlaying images of the same scene taken at different time instants, from different viewpoints, and/or different sensors such that the individual pixels in all images refer to the same physical structure. It is usually accomplished by following four main steps. In what follows we will briefly discuss each one of them [68].

**Feature Detection**

As a first step, salient structures such as closed-boundary regions, edges, contours, line intersections, corners, are detected within the source images. Ideally, these features should be distinct, spread all over the image and efficiently detectable in all source images. In general, feature detection is performed using off-the-shelf solutions such as the Kanade-Lucas-Tomasi feature tracker [69], the scale-invariant feature transform (SIFT) [70] or Harris corners [71], among others.

Please note that there exists a second group of image registration techniques which do not rely on the detection of features. These techniques, coined area-based or direct registration methods exploit common scene characteristics within the input

(a) (b) (c) (d)

**Figure 1.6:** *Illustration of the three different types of linear transforms commonly employed for image registration purposes. (a) Original image. (b) Similarity transform. (c) Affine transform. (d) Projective transform. Original source image taken from [1].*

images and put emphasis on the feature matching step rather than on their detection. For more information on direct registration methods we refer the interested reader to [72].

### Feature Matching

In this step, correspondences between the detected features within the source images are established. Various feature descriptors and similarity measures along with spatial relationships among the features are used for that purpose. After establishing an initial set of correspondences between the extracted feature points, methods such as random sample consensus (RANSAC)[73] may subsequently be employed to refine the matches.

### Transform Model Estimation

After the feature correspondences have been established, the parameters of the transform model need to be estimated. In this context we can differentiate between three types of transforms, namely, similarity transforms, affine transforms and perspective transforms (see Fig. 1.6) [1]. The similarity transform is the simplest model consisting solely of rotation, translation and scaling. It is often called 'shape-preserving mapping' since it preserves angles. A slightly more general model is the affine model which is capable of mapping parallelograms onto squares. It preserves straight line parallelism and is usually used in registration scenarios where the distance of the camera to the scene is large when compared to the field-of-view of the camera. If the condition on the distance of the camera from the scene is not satisfied a projective transform should be used. Projective transforms can map a general quadrangle onto a square and describe the exact deformation of a flat scene photographed by a pinhole camera whose optical axis is not perpendicular to the scene.

Note that the type of mapping function used should correspond to the geometric

deformation assumed between the source images. Furthermore, in order to improve the overall registration accuracy it is often necessary to consider nonlinear distortion effects arising from the optical lens employed in the camera (see Chapter 6 for more details).

**Image Resampling and Transformation**

The mapping function estimated during the previous step is used to transform the source images, thereby registering the misaligned input images. Image values at non-integer coordinate positions are computed by employing an appropriate interpolation technique. In this context, spline functions [74], Gaussians [75] and truncated sinc functions [76] belong to the most commonly used interpolants. ■

In the remainder of this work we assume that all images are adequately aligned and registered prior to the fusion process. The used images were acquired from the internet and represent the same set of images commonly employed by the image fusion community. However, due to the inherent difficulties in producing registered source images at sub-pixel accuracy, they are small in number. Based on this observation, Chapter 6 introduces a novel IR/visible-light image registration technique which is able to register spatially and temporally misaligned image sequences. Thereby, the entire work flow, starting from image acquisition and ending with image fusion, is covered in the presented work.

## 1.5 Outline

Chapter 2 gives a more detailed view on multiscale image fusion. In particular, we discuss the most relevant work reported within the last decades and present a generic framework which encompasses the most important aspects of multiscale image fusion.

In Chapter 3 a performance analysis and performance comparison of several multiscale transforms for the purpose of image fusion is held, representing the first contribution of this work. We start off by conducting a theoretical review on orthogonal and redundant multiscale decompositions such as the DWT, CVT, ConT, UWT, DTCWT and NTSC. Subsequently, the suitability of these transforms for image fusion is investigated using a generic fusion rule. The chapter is concluded with an analysis of the obtained results.

Another contribution of this work is presented in Chapter 4. Based on the conclusions drawn in Chapter 3, we introduce a novel pixel-level UWT-based multiscale fusion framework which uses spectral factorization of the analysis filter pair in combination with non-orthogonal filter banks. We show that this approach is able to

improve the fusion results compared to traditional approaches for a large class of input images.

For IR/visible-light image pairs, fusion results can be further improved by including information about the presence of targets within the IR image into the fusion process. For this purpose, Chapter 5 introduces a novel IR segmentation method which ensures that all identified targets are properly incorporated in the fused image. Additionally, a new hybrid fusion scheme is proposed which utilizes both pixel-level and region-level information to fuse the source images.

Most conclusions drawn in the course of this work are based on the results obtained for a comparatively small set of input image pairs. Thus, Chapter 6 describes the creation of an exhaustive IR/visible-light image data base suitable for image fusion purposes. In this context, a novel IR/visible-light video registration framework is introduced which significantly improves the registration results when compared to the state-of-the-art. By means of the proposed methodology 30 different IR/visible-light video sequences, recorded at 6 different locations, were generated.

Finally, our conclusions are given in Chapter 7.

# Chapter 2

# Multiscale pixel-level fusion

In the last two decades pixel-level image fusion gained considerable attention from the image processing community. The simplest pixel-level image fusion scheme is to take the pixel-by-pixel average of the source images. Such a scheme is presented in Fig. 2.1, where a medical image pair, illustrated in Figs. 2.1(a) and 2.1(b), is fused by averaging. Although the averaging method is very simple to implement, it presents several drawbacks including reduced contrast, which can lead to a severe loss of information, as depicted in Fig. 2.1(c). For comparison purposes, Fig. 2.1(d) represents the fusion result obtained by applying a novel fusion approach which is described in detail in Chapter 4.

Most image fusion approaches operating on pixel-level rely on transform domain techniques to properly combine the source images. While many such transforms have been proposed for image fusion purposes (see Section 1.3.1 for an overview), most transform domain approaches use multiscale decompositions. This is motivated by the fact that images tend to present features in many different scales. In addition, the human visual system seems to exhibit high similarities with the properties of multiscale transforms. More precisely, strong evidence exists that the entire human visual field is covered by neurons that are selective to a limited range of orientations and spatial frequencies, and can detect local features like edges and lines. This makes the neurons response very similar to the basis functions of multiscale transforms [77].

The basic idea underlying multiscale image fusion is to perform a multiscale transform on each source image and, following some specific fusion rules, construct a single composite multiscale representation from these. The final fused image is obtained by taking the inverse transform of the composite representation. This process is illustrated in Fig. 1.5 in Chapter 1 for two input images $I_1$ and $I_2$, where $\mathcal{T}\{\cdot\}$ and $\mathcal{T}^{-1}\{\cdot\}$ represent the forward and inverse tranform, respectively.

In the literature plenty of pixel-level multiscale image fusion works can be found. In what follows, we review some of these approaches and present a generic multiscale pixel-level framework which is able to incorporate most of them. This framework

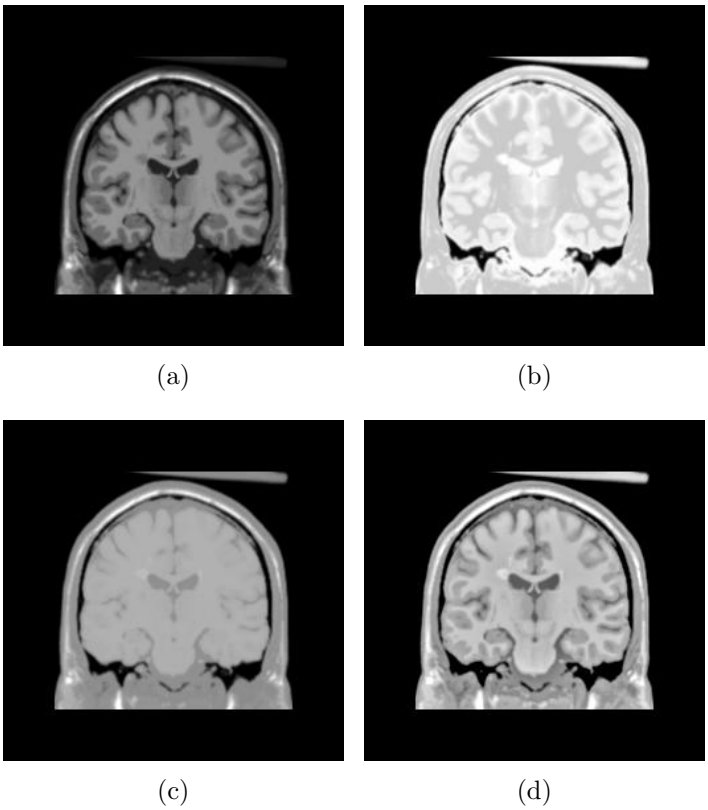


**Figure 2.1:** *Average fusion vs. multiscale fusion. (a) and (b) Medical source image pair. (c) Fusion result by averaging. (d) Fusion result by applying the novel multiscale fusion framework of Chapter 4.*

was first introduced by Piella in [7] and can be seen as an extension of the multiscale methodology proposed by Zhang and Blum in [5].

## 2.1 Notation

Let us start by fixing some notation which is needed in the remainder of this work. When using multiscale transforms, an input image can be represented in the transform domain by a sequence of detail images along with an approximation image at the coarsest scale. Henceforth, the multiscale decomposition of an input image $I_k$ is represented as

$$y_k = \{y_k^1, y_k^2, \ldots, y_k^J, x_k^J\}, \tag{2.1}$$

where $x_k^J$ represents the approximation image at the lowest scale $J$ and $y_k^j, j = 1, \ldots, J$ represent the detail images at level $j$. The detail images comprise in general of various orientation bands, depending on the multiscale transform in use. We assume henceforth that a detail image $y_k^j$ at level $j$ is composed of $P$ detail images $y_k^j = \{y_k^j[\cdot\,, 1], y_k^j[\cdot\,, 2], \ldots, y_k^j[\cdot\,, P]\}$. As in the previous chapter, we let $m, n$ index the location of the pixel or coefficient. When convenient, we also use the vector

$\mathbf{n} = [m, n]$ to specify the spatial location. In such a case $y_k^j[\mathbf{n}, p]$ represents the detail coefficient of the $k^{\text{th}}$ input image at location $\mathbf{n}$, decomposition level $j$ and orientation band $p$. In order to simplify the discussion, we assume, without loss of generality, that the fused image $I_F$ is generated from two source images, $I_A$ and $I_B$. However, most presented techniques are easily expandable to an arbitrary number of input images.

## 2.2 Overview of some existing approaches

The usage of multiscale image transforms is not a recent approach in image fusion applications. The first multiscale image fusion approach was proposed by Burt [37] in 1984 and is based on the Laplacian Pyramid. As for the fusion rule, a simple pixel-based maximum selection rule was used. Thus, each composite coefficient is obtained by

$$y_F^j[m, n] = \begin{cases} y_A^j[m, n] & \text{if } \left|y_A^j[m, n]\right| > \left|y_B^j[m, n]\right| \\ y_B^j[m, n] & \text{otherwise} \end{cases}. \tag{2.2}$$

Motivated by the fact that the human eye is more sensitive to contrast changes than to absolute luminance differences, Toet [78] presented a similar algorithm using the Ratio-of-Low-Pass Pyramid instead of the Laplacian Pyramid.

Burt and Kolczynski [32] proposed the use of the Gradient Pyramid for the fusion of multisensor, multiexposure and multifocus images. In their approach two measures are used to guide the fusion process. The first one is an activity measure $a_k^j$ which is in charge of determining the saliency of the source images at each coefficient position $\mathbf{n}$. It is defined as a local energy measure such that

$$a_k^j[\mathbf{n}, p] = \sum_{\Delta\mathbf{n} \in \mathcal{W}} \left|y_k^j[\mathbf{n} + \Delta\mathbf{n}, p]\right|^2, \tag{2.3}$$

with $\mathcal{W}$ representing a window of size $1 \times 1$, $3 \times 3$ or $5 \times 5$ centered at the origin. The second one is a match measure $m_{AB}^j$ which is used to quantify the similarity between the two pyramid-transformed, input images. It is given by

$$m_{AB}^j[\mathbf{n}, p] = \frac{2 \sum_{\Delta\mathbf{n} \in \mathcal{W}} y_A^j[\mathbf{n} + \Delta\mathbf{n}, p] y_B^j[\mathbf{n} + \Delta\mathbf{n}, p]}{a_A^j[\mathbf{n}, p] + a_B^j[\mathbf{n}, p]} \tag{2.4}$$

and corresponds to a local correlation function. Again, the window $\mathcal{W}$ may include only a single coefficient or a small local area. The actual fusion is defined as a weighted average where at each coefficient position, weights $w_A^j$ and $w_B^j$ are assigned to the transformed, source images. Thus, the fused image in the pyramid transform

domain is defined as

$$y_F^j[\mathbf{n},p] = w_A^j[\mathbf{n},p]y_A^j[\mathbf{n},p] + w_B^j[\mathbf{n},p]y_B^j[\mathbf{n},p], \tag{2.5}$$

where the weights are determined by

$$w_A^j[\mathbf{n},p] = \begin{cases} 1 & \text{if } m_{AB}^j[\mathbf{n},p] \leq T \text{ and } a_k^A[\mathbf{n},p] > a_k^B[\mathbf{n},p] \\ 0 & \text{if } m_{AB}^j[\mathbf{n},p] \leq T \text{ and } a_k^A[\mathbf{n},p] \leq a_k^B[\mathbf{n},p] \\ \frac{1}{2} + \frac{1}{2}\left(\frac{1-m_{AB}^j[\mathbf{n},p]}{1-T}\right) & \text{if } m_{AB}^j[\mathbf{n},p] > T \text{ and } a_k^A[\mathbf{n},p] > a_k^B[\mathbf{n},p] \\ \frac{1}{2} - \frac{1}{2}\left(\frac{1-m_{AB}^j[\mathbf{n},p]}{1-T}\right) & \text{if } m_{AB}^j[\mathbf{n},p] > T \text{ and } a_k^A[\mathbf{n},p] \leq a_k^B[\mathbf{n},p] \end{cases} \tag{2.6a}$$

$$w_B^j[\mathbf{n},p] = 1 - w_A^j[\mathbf{n},p] \tag{2.6b}$$

for some threshold $T$. Observe that in case of a low similarity between the input images (the match measure is below or equal the threshold $T$), the weights are either 1 or 0 which corresponds to a maximum selection rule such as depicted in eq. (2.2). On the other hand, if the similarity is high (match measure is above the threshold $T$) a weighted sum of the coefficients is used.

The use of the Discrete Wavelet Transform (DWT) in image fusion was proposed by Li et al. [40]. In their implementation the maximum absolute value within a window of size $3 \times 3$ or $5 \times 5$ is used as an activity measure and associated to the pixel centered in the window $\mathcal{W}$ such that

$$a_k^j[\mathbf{n},p] = \max_{\Delta\mathbf{n}\in\mathcal{W}} \left|y_k^j[\mathbf{n}+\Delta\mathbf{n},p]\right|. \tag{2.7}$$

A binary decision map is then created to record the selection results based on a maximum selection rule. This decision map is then subject to a consistency verification. Specifically, if the center pixel value comes from $I_A$ while the majority of the surrounding pixel values come from $I_B$, the center pixel is switched to that of $I_B$.

In [79] the authors present a fusion approach which makes use of a steerable dyadic wavelet transform. In their approach, first the local oriented energy is obtained from a quadrature pair of steerable filters and the local dominant orientation (the angle that maximizes the local oriented energy) is computed at each level and position. Based on these calculations, the filters are then steered to the local dominant orientation and the local oriented energies of the input images are compared. The fusion is performed by transferring those coefficients to the composite representation which correspond to the greater local energy. Finally, the filters are steered back to their original orientation and reconstruction is carried out. Liu et al. [38] present another approach based on a steerable dyadic wavelet transform. However,

rather than using it to fuse the source images, they fuse the various bands of the decomposition by means of a Laplacian pyramid.

To overcome the shift dependency of the DWT fusion schemes, Rockinger proposed the use of the Undecimated Wavelet Transform (UWT) instead [44, 57]. The fusion of the detail coefficients was accomplished using two selection schemes: i) a point-based maximum selection rule and ii) an area-based selection scheme with window size $5 \times 5$ and subsequent consistency verification as proposed by Li et al. [40]. The coarse approximation coefficients were combined by a simple averaging operation. The author shows that this approach is particularly useful for image sequence fusion, where a composite image sequence has to be built from various input image sequences.

In 1999, Zhang and Blum [5] tried to map all previously published multiscale image fusion proposals into one generic framework. In this framework, after application of the multiscale transform, an activity-level measure is computed which attempts to determine the quality of each source image. This is followed by an optional grouping of coefficients belonging to the same spatial position in the source image. Based on the activity measure and the coefficient grouping, the composite multiscale representation is obtained using some fusion rule. A consistency verification procedure is then performed which incorporates the idea that a composite coefficient is unlikely to be generated in a completely different manner from all its neighbors. In order to asses the different alternatives for each step of the generic framework, a performance comparison is conducted. The authors conclude that coefficient grouping and consistency verification generally improve the final fusion result, whereas the proper choice of activity measurement and combination method depend to a high extend on the underlying application. As for the multiscale transform, the authors claim that the UWT provides better performance than the DWT and pyramid-based transforms.

Pu and Ni [39] proposed a contrast-based image fusion method using the DWT. For this purpose they introduced an activity measure such that

$$a_k^j[m,n,p] = \left| \frac{y_k^j[m,n,p]}{x_k^J[m,n]} \right|, \tag{2.8}$$

which they called directive contrast and use a maximum selection rule to combine the wavelet coefficients.

In [45] an alternative implementation of the UWT is used for the fusion of multisensor images. Since the filters do not need to be (bi)orthogonal (as opposed to the DWT), they proposed the use of a filter bank where the wavelet coefficients can

be obtained by a simple difference between two successive approximations

$$y_k^{j+1}[m,n] = x_k^j[m,n] - x_k^{j+1}[m,n]. \tag{2.9}$$

After decomposing both source images, the activity measure given in eq. (2.3) is calculated for all detail coefficients. Following the maximum selection fusion rule, the detail coefficient yielding a higher activity is directly transferred to the decomposed, composite image whereas the approximation images are fused by simple averaging. The reconstructed, fused image is obtained by a simple co-addition of all detail images to the approximation image by

$$I_F[m,n] = x_F^J[m,n] + \sum_{j=1}^{J} y_F^j[m,n]. \tag{2.10}$$

Note that, in this approach, for each scale only one detail image is produced and not three as in the general two-dimensional wavelet case.

Another DWT image fusion approach is introduced by Petrović and Xydeas in [41]. In the proposed methodology, the input images are represented at each scale through gradient maps which express the information contained in these images as changes in pixel values, rather than absolute gray-level values. These gradient maps are subsequently fused using a so-called "cross-band pyramid fusion method" [80] where fusion decisions are taken jointly for all coefficients representing the same spatial position in the horizontal, vertical and diagonal directional detail images. Furthermore, since it is assumed that there exists a strong linkage between coefficients in neighboring scales, inter-scale relationships between coefficients belonging to the same spatial position are considered as well. The activity is therefore given by

$$a_k^j[m,n,p] = \sum_{q=1}^{3} \left| y_k^j[m,n,q] \right| + \left| y_k^{j+1}[u,v,q] \right|, \tag{2.11}$$

where $q = 1,2,3$ specifies the orientation band and $u$ and $v$ are defined as $\left\lceil \frac{m}{2} \right\rceil$ and $\left\lceil \frac{n}{2} \right\rceil$, respectively, due to the downsampling involved in every decomposition step. Next, a similarity measure is calculated

$$m_{AB}^j[m,n,p] = \left| \frac{\sum_{q=1}^{3} \left| y_A^j[m,n,q] \right| - \sum_{q=1}^{3} \left| y_B^j[m,n,q] \right|}{\max\left( \sum_{q=1}^{3} \left| y_A^j[m,n,q] \right|, \sum_{q=1}^{3} \left| y_B^j[m,n,q] \right| \right)} \right| \tag{2.12}$$

which, in combination with the activity measure, is used to determine the fusion

weights:

$$\begin{bmatrix} w_A^j[\mathbf{n},p] \\ w_B^j[\mathbf{n},p] \end{bmatrix}^T = \begin{cases} [0.5, 0.5] & \text{if } m_{AB}^j[\mathbf{n},p] < T_1 \\ [1,0] & \text{if } T_1 < m_{AB}^j[\mathbf{n},p] < T_2 \text{ and } a_A^j[\mathbf{n},p] \geq a_B^j[\mathbf{n},p] \\ [0,1] & \text{if } T_1 < m_{AB}^j[\mathbf{n},p] < T_2 \text{ and } a_A^j[\mathbf{n},p] < a_B^j[\mathbf{n},p] \\ [1,1] & \text{if } m_{AB}^j[\mathbf{n},p] > T_2 \end{cases}. \quad (2.13)$$

The fused gradient maps are then computed by using a weighted averaging as given in eq. (2.5). The proposed weight calculation is in spirit very similar to eq. (2.6) since a high similarity (match measure is below the threshold $T_1$) leads to the averaging of the input coefficients whereas in case of low similarity (match measure is between thresholds $T_1$ and $T_2$), a maximum selection rule is used. However, the authors consider a further case where the fused coefficients are evaluated as the sum of both inputs. They motivate this choice by claiming that in situations where the input coefficients are substantially different ($m_{AB}^j > T_2$) both source images exhibit independent features which need to be conserved. Please note that after calculating the composite gradient maps, they are further decomposed into a simplified version of the DWT. Since the actual fusion is performed on the gradient maps rather than in the DWT domain, the authors referred to their contribution as a "fuse-then-decompose" technique.

At the same time Forster et al. [43] proposed the use of complex wavelets for the fusion of multichannel microscopy images. First, the multichannel (color) data is converted to a single channel by an adaptive weighted linear combination of all input channels using a PCA-like approach. Next, a complex-valued discrete wavelet transform is applied on the converted, gray-scale images and the fusion is performed in the transform domain. For this purpose, the largest absolute value of the wavelet coefficients at each point is used (maximum selection rule) and a coefficient grouping and consistency verification step is performed. Finally, the corresponding inverse complex-valued DWT is applied and the fused, gray-scale image is reconverted to obtain multichannel data. This reconversion is accomplished by utilizing the weights computed during the multi-to-single-channel conversion. The authors show that the added phase information of complex wavelets yields stability and better preserves image details during the fusion step, compared to real-valued DWTs.

In order to overcome the shift-variance and limited directionality of the DWT while maintaining the perfect reconstruction property with limited redundancy, Ray and Adhami [46] used the DTCWT for multifocus and multisensor image fusion. In the proposed methodology, the actual fusion is performed by selecting the coefficient with the largest magnitude in the transform domain. However, since the DTCWT is implemented using two filter bank trees, with the first tree representing the real

part and the second tree representing the imaginary part, the coefficient magnitudes are calculated as

$$a_k^j[m,n,p] = \sqrt{y_{real,k}^j[m,n,p]^2 + y_{imag,k}^j[m,n,p]^2}, \tag{2.14}$$

with $y_{real,k}^j$ and $y_{imag,k}^j$ representing the coefficients of the real and imaginary trees, respectively, of the $k^{\text{th}}$ source image.

A multifocus image fusion algorithm which combines the DWT and the Curvelet Transform (CVT) is introduced by Li and Yang [48]. In their approach the source images are first decomposed using the CVT. The resulting detail and approximation images form the input image set for the DWT, which is applied to all CVT-decomposed images subsequently. Finally, the fusion is employed in the wavelet domain by using a maximum selection rule in combination with a consistency verification step. In order to obtain the final fused image, first, the inverse DWT is applied followed by the inverse CVT. The authors motivate this approach with the complementary properties of the two transforms. Whereas the DWT is very efficient in representing isotropic elements such as textures (detail information), the CVT is suitable for catching the edges in an image (structural information). Thus, by combining the two transforms, the authors claim that the fusion process considers both detail and structural information, leading to a fused image which is superior than the result of applying any of the two transforms individually.

In [51] another fusion framework is proposed which attempts to avoid the shortcomings of wavelet-based methods by using the Nonsubsampled Contourlet Transform (NSCT) for the fusion of multifocus images. Like the CVT and the Contourlet Transform (ConT), the NSCT is a further representative of a new family of transforms which possess anisotropic basis elements and can be implemented using an (almost) arbitrary number of directions at each scale. As for the fusion of the approximation images, in the presented work a so-called "clarity measure" is utilized that measures the activity of the detail images at the coarsest scale in the NSCT-domain such that

$$a_k^J[m,n] = \sqrt{\sum_{p=1}^{P} y_k^J[m,n,p]^2}. \tag{2.15}$$

The composite approximation image is then defined as

$$x_F^J[m,n] = \begin{cases} x_A^J[m,n] & \text{if } a_A^J[m,n] - a_B^J[m,n] > T \\ \dfrac{x_A^J[m,n] + x_B^J[m,n]}{2} & \text{if } \left|a_A^J[m,n] - a_B^J[m,n]\right| < T\,, \\ x_B^J[m,n] & \text{if } a_B^J[m,n] - a_A^J[m,n] > T \end{cases} \tag{2.16}$$

where $T$ is an experimentally obtained threshold. For the combination of the detail

images, first, an activity measure is defined which uses the directive contrast of eq. (2.8) in combination with the standard deviation of the set of detail coefficients representing the same spatial position at the same scale. Hence, the activity measure for the detail images in the NSCT-domain is given by

$$a_k^j[m,n,p] = \left| \frac{y_k^j[m,n,p]}{x_k^J[m,n]} \right| \sqrt{\frac{1}{Q} \sum_{q=1}^{Q} \left( y_k^j[m,n,q] - \bar{y}_k^j[m,n] \right)^2}, \tag{2.17}$$

where $q = 1, \ldots, Q$ specifies the orientation band and

$$\bar{y}_k^j[m,n] = \frac{1}{Q} \sum_{q=1}^{Q} \left| y_k^j[m,n,q] \right|. \tag{2.18}$$

The authors argue that by this choice, the influence of noise is minimized since its energy distribution is uniformly spread over all directions, leading to a small standard deviation. The combination of the detail coefficients is finally accomplished by a maximum selection rule which transfers the coefficient with higher activity directly to the fused NSCT decomposition.

More recently, Li et al. [2] conducted a performance study on different multiscale transforms for image fusion. They concluded that the best results for medical, multifocus and multisensor image fusion can be achieved using the NSCT, followed by the DTCWT and the UWT. In their experiments they utilized the maximum selection rule for all detail coefficients whereas a simple averaging operation was applied to the approximation images. Additionally, recommendations regarding filter choices and number of directional decompositions, where applicable, are given.

## 2.3 A generic multiscale pixel-level image fusion framework

Inspired by the fusion methodology proposed by Zhang and Blum in [5], Fig. 2.2 shows a generic multiscale pixel-level image fusion framework which is able to encompass most fusion schemes discussed so far. It was first introduced by Piella [7] and can be seen as a more detailed version of Fig. 1.5 in which the combination algorithm is elaborated more carefully. In the remainder of this section, we describe the individual building blocks in more detail.

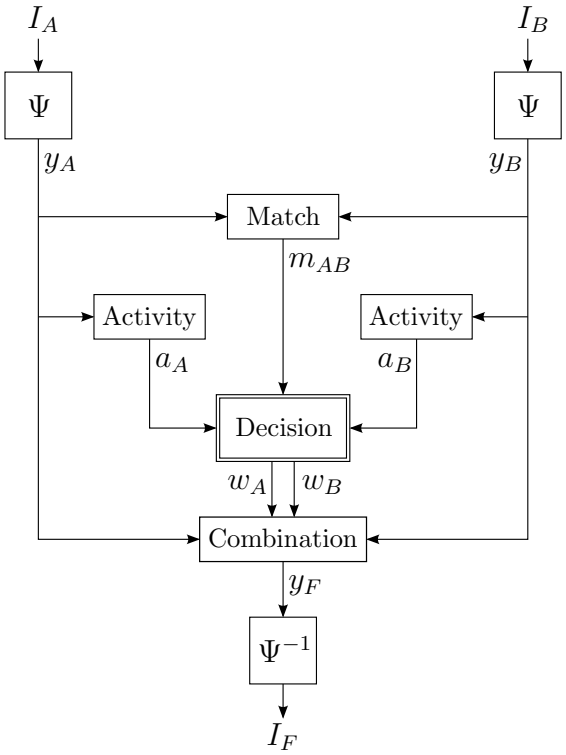



**Figure 2.2:** *Generic multiscale pixel-level fusion framework.*

### 2.3.1 Multiscale Decomposition ($\Psi$)

As we have already seen before, a multiscale transform decomposes a given input signal into a set of signals which comprises information at different scales. In such a representation the high levels contain low-frequency information while low levels contain the high-frequency information. Furthermore, multiscale representations are generally able to reveal the information we are looking for in a few coefficients which allows for a very efficient representation of the underlying input data [81]. The use of multiscale transforms is suitable for image fusion, not only because it enables us to consider and fuse image features separately at different scales, but also because such a sparse representation only produces large coefficients near important image structures like edges, thus revealing salient information. A large part of research on multiscale image fusion has been focused on choosing an appropriate transform. Apart from the fact that the final decision is highly application dependent, the main issues addressed in this respect are:

**Shift-invariance**

As stated in various studies (e.g. [2], [5] and [23]), shift-invariance, as provided by the UWT, NSCT and approximately by the DTCWT is a highly desirable property in image fusion applications. Shift-dependency is especially problematic considering misregistration problems and in image sequence fusion. However, shift-invariant transforms often come with the handicap of an increased degree of redundancy.

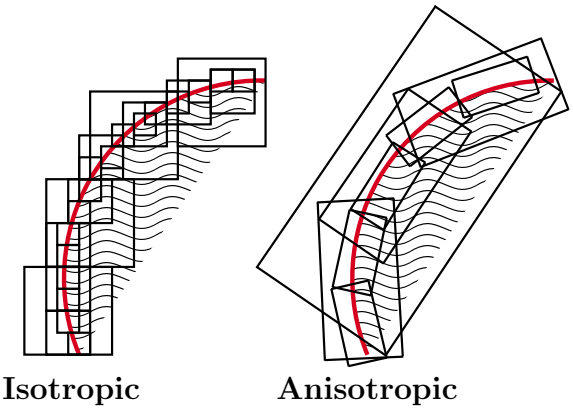



**Figure 2.3:** *Representation of a curve, separating two smooth regions using isotropic and anistropic basis elements.*

### Orthogonality

In general, redundant transforms lead to better fusion results than orthogonal transforms such as the DWT. This is mainly due to the over-complete set of basis functions, provided by redundant transforms, which are able to capture the intrinsic properties of images better than orthogonal decompositions [81]. Furthermore, the sampling involved in orthogonal transforms often causes a deterioration in the quality of the fused image by introducing heavier blocking effects [7]. However, redundant transforms generally lead to an increase in data volume and complexity which may limit their use in some situations. Note that shift-invariance and redundancy often go hand in hand.

### Anisotropy

To turn the feature selection process more robust and minimize the introduction of distortions in the fused image, it is advantageous to represent salient information in the source images with as few coefficients as possible. In natural images such information is typically exhibited by discontinuity points such as edges which are located along curves belonging to the boundaries of physical objects [82]. Traditional transforms such as the DWT, UWT and most pyramid-based transforms, however, fail to efficiently represent these structures since they rely on a dictionary of roughly isotropic basis elements. In other words, these transforms do not "see" the smoothness along the curve and require a significant number of coefficients to represent it. Transforms like the CVT, ConT and the NSCT, on the other hand, use anisotropic basis elements with elongated shape which are able to represent smooth curves more efficiently. Fig. 2.3 illustrates the effect of using isotropic and anisotropic basis elements to represent a smooth contour.

### Directionality

Another important attribute of multiscale transforms is the number of directional

decompositions offered per scale. While most pyramid-based transforms fail do provide any directional information, wavelet transforms exhibit three directional detail images, corresponding to the horizontal, vertical and diagonal direction. By using the CVT, ConT or NSCT an (almost) arbitrary number of directions can be implemented. In general, a higher directional selectivity results in a more compact representation of image features. Usually, there exists a strong connection between anisotropy and multidirectionality. In fact, it is because of these anisotropic basis elements that multidirectionality is possible [82]. ■

Other important aspects which may influence the overall fusion performance are the number of decomposition levels and the used filter bank. In Chapter 3 a performance comparison of several multiscale transforms together with an analysis of the results for different numbers of decomposition levels and filter settings is conducted.

### 2.3.2 Activity measure

The activity measure is in charge of reflecting the degree of saliency exhibited by a single coefficient within a given subband. For example, when combining images having different foci, a desirable activity measure would provide a quantitative value that increases whenever an object is in focus. At pixel-level there exist two classes of methods to compute the activity, namely, coefficient-based and window-based measures [5]. The coefficient-based activity measures consider each coefficient separately. In this case, the activity is usually calculated by taking the absolute value of the coefficient, given by

$$a_k^j[m,n,p] = \left|y_k^j[m,n,p]\right|. \tag{2.19}$$

In contrast, window-based activity measures employ a small (typically $3 \times 3$ or $5 \times 5$) window centered at the current coefficient position. A diagram illustrating these two types of activity measures is shown for a single subband in Fig. 2.4.

Based on the fact that the human visual system is primarily sensitive to local contrast changes, most window-based activity measures employ some sort of energy calculation

$$a_k^j[\mathbf{n},p] = \sum_{\Delta\mathbf{n}\in\mathcal{W}} w_k[\Delta\mathbf{n}] \left|y_k^j[\mathbf{n}+\Delta\mathbf{n},p]\right|^\gamma, \quad \gamma \in \mathbb{R}_+, \tag{2.20}$$

where $w_k$ are the window weights with $\sum_{\Delta\mathbf{n}\in\mathcal{W}} w_k[\Delta\mathbf{n}] = 1$. Alternatively, one can compute the activity as the contrast of a single coefficient with its neighbors

$$a_k^j[\mathbf{n},p] = \frac{\left|y_k^j[\mathbf{n},p]\right|}{\sum_{\Delta\mathbf{n}\in\mathcal{W}} w_k[\Delta\mathbf{n}] \left|y_k^j[\mathbf{n}+\Delta\mathbf{n},p]\right|}. \tag{2.21}$$

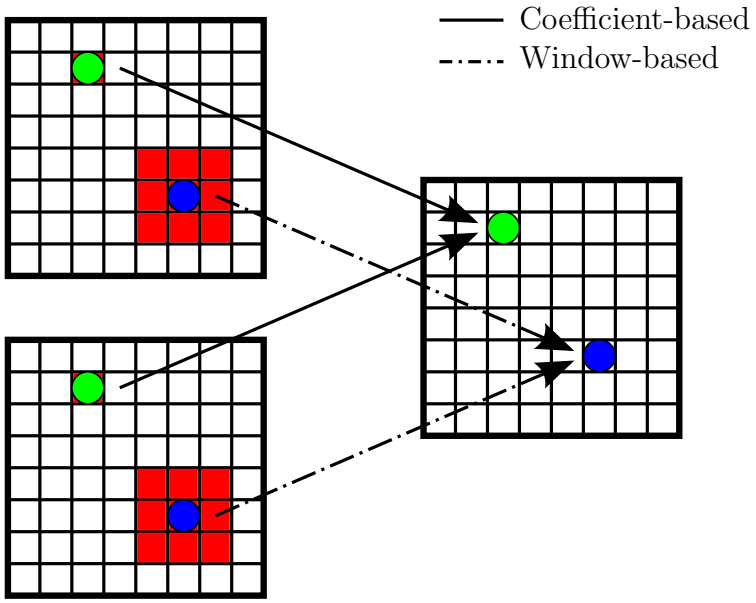



**Figure 2.4:** *Fusion of a single subband using a coefficient-based and a window-based activity measure.*

Another approach is to use a non-linear criteria such as the so-called rank filter which considers the $i^{\text{th}}$ highest value within a small window

$$a_k^j[\mathbf{n}, p] = \underset{\Delta\mathbf{n}\in\mathcal{W}}{\text{rank(i)}} \left| y_k^j[\mathbf{n} + \Delta\mathbf{n}, p] \right| \tag{2.22}$$

as an appropriate activity measure. A special case of the rank filter is the median filter which can be used to e.g. turn the activity measure more robust against impulsive noise.

Finally, we would like to point out that the presented list of activity measures is not exhaustive. For example, one might also consider to calculate the activity by computing the spatial frequency measure as described in [83] or utilize statistical properties such as the mean or the standard deviation within a window.

### 2.3.3 Match measure

The match or similarity measure between the transform coefficients of the source images expresses to which extent the source images differ. In combination with the activity measure, this information is used to reach an appropriate fusion decision. The match measure between $y_A^j$ and $y_B^j$ is usually expressed in terms of a local correlation measure averaged over a neighborhood of the samples

$$m_{AB}^j[\mathbf{n}, p] = \frac{2 \sum\limits_{\Delta\mathbf{n}\in\mathcal{W}} w_k[\Delta\mathbf{n}] y_A^j[\mathbf{n} + \Delta\mathbf{n}, p] y_B^j[\mathbf{n} + \Delta\mathbf{n}, p]}{\sum\limits_{\Delta\mathbf{n}\in\mathcal{W}} w_k[\Delta\mathbf{n}] \left( \left| y_A^j[\mathbf{n} + \Delta\mathbf{n}, p] \right|^2 + \left| y_B^j[\mathbf{n} + \Delta\mathbf{n}, p] \right|^2 \right)}, \tag{2.23}$$

where, again, $w_k$ are the window weights such that $\sum_{\Delta\mathbf{n}\in\mathcal{W}} w_k[\Delta\mathbf{n}] = 1$.

### 2.3.4 Decision method

The decision on how to combine the source images is the key point in most image fusion approaches since it controls the construction of the fused, decomposed image. The output of the decision step is, in general, a set of fusion weights which are stored in a so-called decision map.

Decision mechanisms can be broadly divided into purely selective, purely arithmetic or composite schemes which represent a combination of the first two. As for selective schemes, a natural approach would be to assign a fixed weight of one to the coefficient which exhibits the highest degree of saliency, e.g. the one with the largest activity. This is one of the simplest weighting schemes and is known in the literature as a "choose max" selection or maximum selection rule. For the case of two input sources, the fusion weights are defined as

$$w_A^j[m,n,p] = \begin{cases} 1 & \text{if } a_A^j[m,n,p] > a_B^j[m,n,p] \\ 0 & \text{otherwise} \end{cases} \tag{2.24a}$$

$$w_B^j[m,n,p] = 1 - w_A^j[m,n,p]. \tag{2.24b}$$

Since the sum of the individual weights always has to be one, we henceforth omit the calculation of $w_B^j$ and consider eq. (2.24b) to be valid for all remaining cases. Eq. (2.24) works well under the assumption that at each image location, only one of the source images provides the most useful information. However, like most selective schemes, the maximum selection approach suffers from a low robustness to noise and random selections, resulting in a "salt and pepper" appearance of the decision maps.

Alternatively, we could use an arithmetic method which assigns to each coefficient a weight that depends increasingly on the activity, for example

$$w_A^j[m,n,p] = \frac{a_A^j[m,n,p]}{a_A^j[m,n,p] + a_B^j[m,n,p]}. \tag{2.25}$$

In general, these schemes lead to a stabilization of the decision map but introduce the problem of contrast reduction in the fused image, in case of opposite contrast in the source images.

In order to overcome, to a certain extend, the problems associated with the selective and arithmetic decision schemes, a composite scheme may be utilized. These decision methods usually employ a match measure to decide whether a selective or

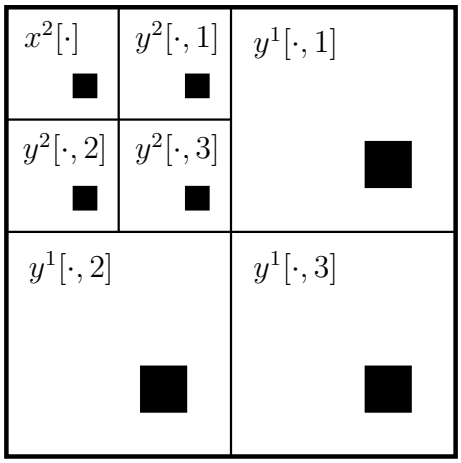



**Figure 2.5:** *Related coefficients in the DWT domain (dark squares), belonging to the same location in the spatial domain.*

arithmetic weight calculation should be performed, for instance,

$$w_A^j[m,n,p] = \begin{cases} 1 & \text{if } m_{AB}^j[m,n,p] \leq T \text{ and } a_A^j[m,n,p] > a_B^j[m,n,p] \\ 0 & \text{if } m_{AB}^j[m,n,p] \leq T \text{ and } a_A^j[m,n,p] \leq a_B^j[m,n,p] \\ \frac{1}{2} & \text{if } m_{AB}^j[m,n,p] > T \end{cases} \tag{2.26}$$

for some threshold $T$. Thus, at locations where the source images are distinctly different, the combination process selects the most salient component, while at locations where they are similar, the average of the source images is taken. In this way, averaging reduces noise and provides stability at locations where source images contain similar information, whereas selection retains salient information and reduces artifacts due to opposite contrast at locations where both source images are different.

In the examples presented so far, the decision is taken for each coefficient independently without reference to the others. However, as illustrated in Fig. 2.5, each coefficient has a set of corresponding coefficients in other directional bands and other decomposition levels which refer to the same spatial location in the source image. Thus, in order to conserve a certain feature from one of the source images in the fused image, all the corresponding coefficients have to be transferred to the composite multiscale representation. Failing to do so may result in a degradation of the fusion result due to the possibility of feature cancellation when the inverse transform is applied. It seems therefore reasonable to consider all (or a subset of) these coefficients when calculating the fusion weights. For example, one may use

the intra-scale dependencies at each decomposition level to obtain

$$w_A^j[m,n,p] = \begin{cases} 1 & \text{if } \sum_{q=1}^{Q} a_A^j[m,n,q] > \sum_{q=1}^{Q} a_B^j[m,n,q] \\ 0 & \text{otherwise} \end{cases}, \tag{2.27}$$

where $q = 1, \ldots, Q$ defines the directional band. In this case, the fusion weights are obtained by a selective rule which takes into account the corresponding activity values of all directional detail images. The most restrictive case is to consider inter-scale dependencies as well, ensuring that all corresponding coefficients receive the same fusion weight. Such a scheme may be given by

$$w_A^j[m,n,p] = \begin{cases} 1 & \text{if } \sum_{l=1}^{L}\sum_{q=1}^{Q} a_A^l[m,n,q] > \sum_{l=1}^{L}\sum_{q=1}^{Q} a_B^l[m,n,q] \\ 0 & \text{otherwise} \end{cases}, \tag{2.28}$$

where $l = 1, \ldots, L$ and $q = 1, \ldots, Q$ represent the decomposition level and directional band, respectively.

Another possibility to improve the overall fusion result is to exploit the fact that it is very likely that a good fusion method computes neighboring coefficients in the composite representation in a similar manner. One example which is based on this idea is the consistency verification method, proposed by Li et al. in [40]. This approach consists of applying a majority filter to a binary decision map, obtained by e.g. a maximum selection rule. For example, consider the case where the center weight within a small window ($3 \times 3$ or $5 \times 5$) indicates that the composite coefficient $y_F^j$ should be selected from $y_A^j$ whereas the majority of the surrounding coefficients should be taken from $y_B^j$. After applying consistency verification, the decision map indicates that the composite coefficient $y_F^j$ should also be selected from $y_B^j$. Note that this mechanism is able to remove the largest amount of selection randomness from the decision map, thus minimizing noise effects. ■

So far in this subsection we made no explicit distinction between approximation and detail images. In fact, all presented techniques so far may also be applied to obtain the fusion weights for the low-pass coefficients. However, because of their different physical meaning, the approximation and detail images are usually treated differently by the decision algorithm. As for the detail images, perceptually important information can be related to the absolute coefficient values. Here, a large value corresponds to sharp intensity changes and, thus, to salient features in the image such as edges, lines or other discontinuities. The nature of the approximation image, however, is different. It represents a coarse representation of the source image

and may exhibit some of its properties such as the main intensity or some coarse textural information. Thus, approximation coefficients with a large absolute value do not necessarily correspond to important features within the source image.

In many approaches, the composite approximation coefficients are obtained by a simple averaging operation. Thus, the approximation fusion weights $v$ are

$$v_A[m,n] = v_B[m,n] = \frac{1}{2}. \tag{2.29}$$

We would like to point out that this approach is based on the assumption that all relevant features have already been captured by the detail images. If this is not valid, one might also consider decision schemes which utilize some activity measurement, based on quantities such as entropy or variance. However, as was pointed out in [41], approximation image fusion methods have little influence on the overall fusion performance. ■

Finally, we want to remark that other factors may also influence the assembling of the decision map. In particular, if some a priori knowledge about the source images is available, the decision block can use such information to further improve fusion performance.

### 2.3.5 Combination method

This block is in charge of performing the actual combination of the transform coefficients of the two source images. This is usually accomplished using a linear mapping

$$y_F^j[m,n,p] = w_A^j[m,n,p]y_A^j[m,n,p] + w_B^j[m,n,p]y_B^j[m,n,p], \tag{2.30a}$$
$$x_F^J[m,n,p] = v_A[m,n]x_A^J[m,n] + v_B[m,n]x_B^J[m,n], \tag{2.30b}$$

where the weights $w_A^j$, $w_B^j$, $v_A$ and $v_B$ are obtained from the decision map. ■

In the next chapter the influence of different multiscale transforms for the purpose of pixel-level image fusion is investigated. We start our discussion with a theoretical review of traditional as well as recently developed image decomposition methods. Next, some fusion results obtained by using each transform with varying decomposition levels and filter banks are presented. Finally, by comparing the achieved fusion results, we give the best candidates for the fusion of three different classes of input images.

# Chapter 3

# Performance comparison of different multiscale transforms for image fusion

Multiscale transforms are among the most popular techniques in image fusion applications. The basic methodology underlying these approaches is to first decompose the source images into a set of images at different scales and orientations before combining each subband image individually in the transform domain. The final fused image is obtained by applying the inverse transform to the composite multiscale representation. This process is illustrated in Fig. 1.5 of Chapter 1 for the case of two input images where $\mathcal{T}$ and $\mathcal{T}^{-1}$ represent the forward and inverse transform, respectively.

The main reason of performing the fusion in the transform domain is that salient features within the source images, such as edges, lines or other discontinuities, result in high coefficient values and are therefore more clearly depicted than in the spatial domain. In addition, strong evidence exists that the human visual system exhibits high similarities with the properties of multiscale transforms, further motivating its use for the purpose of image fusion.

Plenty of multiscale transforms have been proposed in the context of image fusion (see Section 2.2 for a general overview). They range from traditional transforms such as the Discrete Wavelet Transform (DWT) and the Undecimated Wavelet Transform (UWT) to recently developed decompositions like the Dual-Tree Complex Wavelet Transform (DTCWT), the Curvelet Transform (CVT), the Contourlet Transform (ConT) and the Nonsubsampled Contourlet Transform (NSCT).

In the literature only little research effort can be found which attempts to assess the fusion performance of these transforms. For this purpose, a detailed performance comparison of multiscale transforms in the context of image fusion is conducted in

this chapter. We start off by giving a theoretical background on each decomposition before investigating its suitability for image fusion using the generic "choose max" fusion rule for all detail images in combination with a simple averaging of the approximation images. In addition, each transform is tested for a varying number of decomposition levels and different filter banks, thus permitting us to better understand the influence of these settings on the final fusion result. Finally, we conclude the chapter with an analysis of the obtained results.

## 3.1 Multiscale transforms

In this section we briefly review the theory behind some of the most utilized multiscale transform in the context of image fusion. For the sake of consistency with the subsequent chapters we start our discussion with the DWT, CVT and ConT before moving on to shift-invariant transforms such as the UWT, DTCWT and NSCT. Where applicable, the same notation as described in Section 2.1 is used.

### 3.1.1 Discrete Wavelet Transform

With the advent of wavelet theory in the last decade, multiscale methods have become increasingly popular within the signal processing community, with applications ranging from quantum physics to signal coding [84]. In what follows, we give a brief introduction to the DWT emphasizing its relation to filter banks. More detailed background texts analyzing the DWT from other points of view can be found in e.g. [81], [85], [86] and [87].

In a nutshell, the DWT replaces the infinitely oscillating sinusoidal basis functions of the Fourier transform with a set of locally oscillating basis functions called wavelets. In the classical setting, wavelets are stretched and shifted versions of a fundamental, real-valued bandpass function $\psi(t)$. When carefully chosen and combined with shifts of a real-valued low-pass scaling function $\phi(t)$, the discrete approximation $x^0$ of an one-dimensional (1-D) finite-energy analog signal $x$ can be decomposed in terms of wavelet and scaling functions via [85]

$$x^0[n] = \sum_{l=-\infty}^{\infty} x^J[l]\tilde{\phi}_{J,l}(n) + \sum_{l=-\infty}^{\infty}\sum_{j=1}^{J} y^j[l]\tilde{\psi}_{j,l}(n), \tag{3.1}$$

with $\tilde{\phi}_{j,l}(t) = 2^{-j/2}\tilde{\phi}(2^{-j}t-l)$ and $\tilde{\psi}_{j,l}(t) = 2^{-j/2}\tilde{\psi}(2^{-j}t-l)$. The approximation (or scaling) coefficients $x^J$ and the detail (or wavelet) coefficients $y^j$ at scale $j = 1, \ldots, J$

are computed via the inner products

$$x^J[n] = \langle x^0, \phi_{J,l} \rangle = \sum_{l=-\infty}^{\infty} x^0[l]\phi_{J,l}(n), \tag{3.2}$$

$$y^j[n] = \langle x^0, \psi_{j,l} \rangle = \sum_{l=-\infty}^{\infty} x^0[l]\psi_{j,l}(n), \tag{3.3}$$

where the family of functions $\{\phi_{J,l}\}_{l\in\mathbb{Z}}$ and $\{\psi_{j,l}\}_{j,l\in\mathbb{Z}}$ are the duals of (orthogonal to) $\{\tilde{\phi}_{J,l}\}_{l\in\mathbb{Z}}$ and $\{\tilde{\psi}_{j,l}\}_{j,l\in\mathbb{Z}}$, respectively. For the particular case where $\phi = \tilde{\phi}$ and $\psi = \tilde{\psi}$ the scaling and wavelet functions form a orthonormal basis expansion. Note that the scaling coefficients $x^J$ and wavelet coefficients $y^j$ provide a time-frequency analysis of the signal by measuring its frequency content (controlled by the scale factor $j$) at different sampling positions (controlled by the spatial shift $l$). This is conceptually very close to the windowed or short-time Fourier transform where a signal is also decomposed using a window which is localized in both the spatial and frequency domain. However, as opposed to the windowed Fourier transform, the DWT uses a time-frequency resolution that changes.

There exists a very efficient way to obtain the wavelet decomposition of the signal $x^0$ without explicitly computing eqs. (3.1), (3.2) and (3.3). More specifically, it was shown in [81] that the scaling and wavelet coefficients $x^J$ and $y^j$ can be related to a two-channel perfect reconstruction filter bank by

$$x^{j+1}[n] = \sum_{l=-\infty}^{\infty} h[l-2n]x^j[l] = \left(\bar{h} * x^j\right)[2n], \tag{3.4}$$

$$y^{j+1}[n] = \sum_{l=-\infty}^{\infty} g[l-2n]x^j[l] = \left(\bar{g} * x^j\right)[2n], \tag{3.5}$$

where $h$ and $g$ are the deployed low-pass and high-pass analysis filters, respectively, and $\bar{h}[n] = h[-n]$. The reconstruction at scale $j$ is obtained by upsampling the coefficients $x^{j+1}$ and $y^{j+1}$, filtering with the synthesis filters $\tilde{h}$ and $\tilde{g}$ and summing the respective outputs. This is given by

$$\begin{aligned} x^j[n] &= \sum_{l=-\infty}^{\infty} \tilde{h}[n-2l]x^{j+1}[l] + \sum_{l=-\infty}^{\infty} \tilde{g}[n-2l]y^{j+1}[l] \\ &= (\tilde{h} * \check{x}^{j+1})[n] + (\tilde{g} * \check{y}^{j+1})[n], \end{aligned} \tag{3.6}$$

where $\check{x}$ denotes the upsampling of $x$ achieved by

$$\check{x}[n] = \begin{cases} x[p] & \text{if } n = 2p \\ 0 & \text{if } n = 2p+1. \end{cases} \tag{3.7}$$

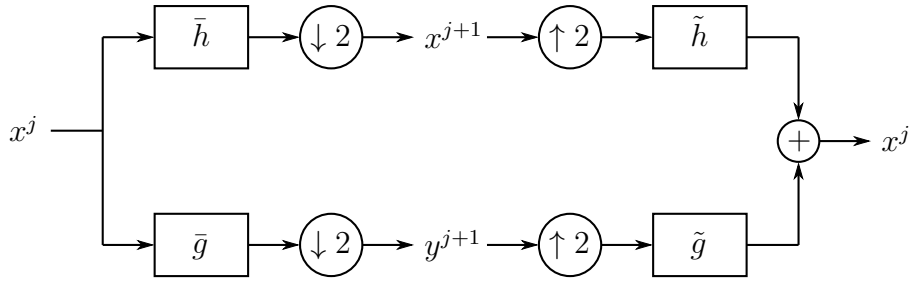



**Figure 3.1:** *DWT two-channel perfect reconstruction filter bank with one decomposition level.*

Fig. 3.1 shows such a two-channel filter bank with a single decomposition level.

For perfect reconstruction, the analysis and synthesis filters need to satisfy the perfect reconstruction condition

$$H(z)\tilde{H}(z) + G(z)\tilde{G}(z) = 1 \tag{3.8}$$

and the anti-aliasing condition

$$H(-z)\tilde{H}(z) + G(-z)\tilde{G}(z) = 0 \tag{3.9}$$

in the $z$-transform domain [81].

The two-dimensional (2-D) DWT decomposition at location $m$,$n$ is given by

$$\begin{aligned}
x^{j+1}[m,n] &= \left(\bar{h}\bar{h} * x^j\right)[2m, 2n] && (3.10)\\
y_1^{j+1}[m,n] &= \left(\bar{h}\bar{g} * x^j\right)[2m, 2n] && (3.11)\\
y_2^{j+1}[m,n] &= \left(\bar{g}\bar{h} * x^j\right)[2m, 2n] && (3.12)\\
y_3^{j+1}[m,n] &= \left(\bar{g}\bar{g} * x^j\right)[2m, 2n] && (3.13)
\end{aligned}$$

where the rows and columns are filtered separately by $h$ and $g$, leading to three high-pass (or detail) images $y_1$, $y_2$, $y_3$ per stage, corresponding to the horizontal, vertical and diagonal directions. The approximation (or low-pass) image $x^j$ is recovered from the coarser-scale approximation image $x^{j+1}$ and the detail images $y^{j+1}$ by two-dimensional separable convolutions in a similar way than in the 1-D case

$$\begin{aligned}
x^j[m,n] &= \left(\tilde{h}\tilde{h} * \check{x}^{j+1}\right)[m,n] + \left(\tilde{h}\tilde{g} * \check{y}_1^{j+1}\right)[m,n] \\
&\quad + \left(\tilde{g}\tilde{h} * \check{y}_2^{j+1}\right)[m,n] + \left(\tilde{g}\tilde{g} * \check{y}_3^{j+1}\right)[m,n]. \qquad (3.14)
\end{aligned}$$

Fig. 3.2 shows the DWT decomposition of Fig. 1.3(a) using the biorthogonal CDF 5/3 filter bank with 2 decomposition levels. This filter bank is also used in the JPEG-2000 standard [88] for lossless compression.

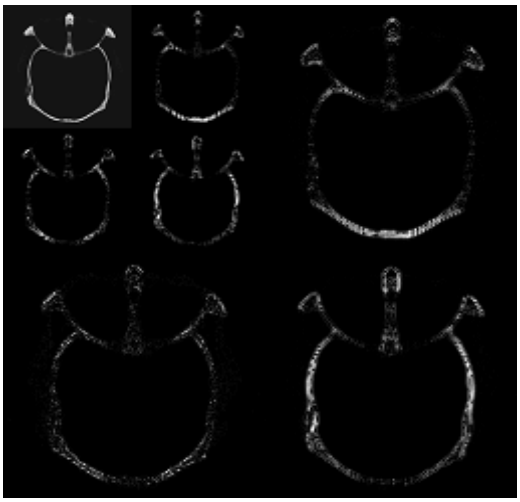


**Figure 3.2:** *DWT decomposition of Fig. 1.3(a) using the biorthogonal CDF 5/3 filter bank with 2 decomposition levels. Each scale and direction has been normalized such that the full dynamic range is occupied.*

### 3.1.2 Curvelet Transform

Despite considerable success of the wavelet theory, research has shown that the classical DWT is far from being universally effective. In image processing for example, one has to deal with the fact that salient information may be located along curves or edges. While wavelets are good at isolating the discontinuities at edge points, they are in general ill-suited for providing a compact representation of such geometric structures. For this purpose, considerable research effort has been put into the development of new transforms which combine ideas from geometry with ideas from traditional multiscale analysis. A special member of this emerging family of multiscale transforms is the Curvelet Transform (CVT) which is briefly introduced in the remainder of this subsection. A more thorough discussion on the CVT can be found in [89] and [90].

In order to better understand the CVT, we start with its continuous implementation and a pair of windows $W(r)$ and $V(s)$ which we call the "radial window" and the "angular window". They are both smooth, non-negative and real-valued, with $W$ taking positive real arguments and being supported on $r \in (\frac{1}{2}, 2)$ and $V$ taking real arguments and being supported on $s \in [-1, 1]$. Now, for each scale $j$ a frequency window $U_j$ is introduced which is defined in the Fourier domain by

$$U_j(r, \theta) = 2^{-3j/4} W(2^{-j} r) V\left(\frac{2^{\lfloor j/2 \rfloor} \theta}{2\pi}\right), \tag{3.15}$$

where $r$ and $\theta$ are polar coordinates. Thus, the support of $U_j$ in the frequency domain is a polar "wedge" defined by the support of $W$ and $V$.

The "mother" curvelet $\phi_j(\mathbf{x})$ for $\mathbf{x} = (x, y)$ is defined by means of its Fourier

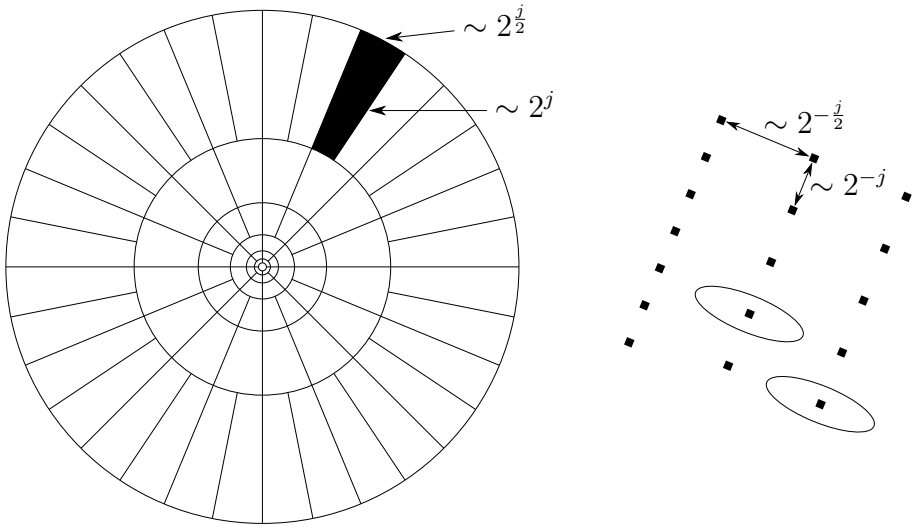



**Figure 3.3:** *Curvelet tiling of space and frequency. The figure on the left represents the induced tiling of the frequency plane. In the frequency domain, curvelets are supported near a "parabolic" wedge, represented by the the highlighted black area. The figure on the right schematically represents the Cartesian grid associated with a given scale and orientation.*

transform $\hat{\phi}_j(\boldsymbol{\omega}) = U_j(\boldsymbol{\omega})$ with $\boldsymbol{\omega} = (\omega_1, \omega_2)$, where we slightly abuse notations by letting $U_j(\omega_1, \omega_2)$ be the window defined in the polar coordinate system by eq. (3.15). Finally, all curvelets at scale $j$ are obtained by rotations and translations of the mother curvelet $\phi_j(\mathbf{x})$. A CVT coefficient $c$ is then simply the inner product between a given function $f \in L^2(\mathbf{R}^2)$ and a curvelet $\phi_{j,k,l}$ such that

$$c(j,k,l) = \langle f, \phi_{j,k,l} \rangle = \int_{\mathbf{R}^2} f(\mathbf{x}) \overline{\phi_{j,k,l}(\mathbf{x})} d\mathbf{x}, \tag{3.16}$$

where $j$, $k$ and $l$ are the scale, rotation and translation parameters, respectively, and $\overline{\phi(\mathbf{x})}$ donates the complex conjugate of $\phi(\mathbf{x})$.

As in wavelet theory, we also have coarse-scale elements which, as opposed to fine-scale elements, are non-directional. At scale $j_0$ these coarse-scale curvelets are defined as

$$\hat{\phi}_{j_0}(\boldsymbol{\omega}) = 2^{-j_0} W_0\left(2^{j_0}|\boldsymbol{\omega}|\right) \tag{3.17}$$

in the frequency domain. Fig. 3.3 summarizes the key components of the CVT construction by showing the curvelet tiling of space and frequency. Note that the effective support length and width of $\phi_j$ obeys the anisotropic scaling relation

$$length \approx 2^{-\frac{j}{2}}, \quad width \approx 2^{-j} \quad \Rightarrow \quad width \approx length^2. \tag{3.18}$$

Thus, the CVT achieves optimal approximation behavior for 2-D piecewise smooth functions that are $C^2$ (twice continuously differentiable functions) except for discontinuities along $C^2$ curves.

To compute the discrete CVT of a digital image $I$ one must take into account the discrete sampling grid, which imposes constraints on the curvelet angles. In a nutshell, the discrete CVT replaces the polar tiling in the frequency domain (see Fig. 3.3) by a discrete rectar-polar tiling. It is implemented by the following steps:

1. Apply the 2-D Fast Fourier Transform (FFT) to the discrete image and obtain Fourier samples $\hat{f}[\omega_1, \omega_2]$

2. For each scale/angle pair $(j, k)$ resample $\hat{f}[\omega_1, \omega_2]$ to obtain sampled values $\hat{f}[\omega_1, \omega_2 - \omega_1 \tan \theta_l]$

3. Multiply the interpolated, Fourier transformed image $\hat{f}$ with the parabolic window $\tilde{U}_j$, localizing $\hat{f}$ near the wedge with orientation $\theta_l$, and obtain

$$\tilde{f}_{j,k}[\omega_1, \omega_2] = \hat{f}[\omega_1, \omega_2 - \omega_1 \tan \theta_k]\tilde{U}_j[\omega_1, \omega_2], \tag{3.19}$$

   where $\tilde{U}_j$ is the Cartesian equivalent to the "polar" window $U_j$ of eq. (3.15).

4. Apply the inverse 2-D FFT to each $\tilde{f}_{j,k}$ hence collecting the discrete curvelet coefficients $c$.

Note that the design of appropriate digital curvelets at the finest scale is not as straightforward as it is for the coarser scales. This is mainly a boundary/periodicity issue. More specifically, the wedge-shaped frequency support of the CVT at finer scales does not fit entirely in the fundamental cell and its periodization may introduce energy at unwanted angles. This problem can be solved by assigning non-directional wavelets to the finest scale.

The discrete implementation of the CVT does not represent a critically-sampled transform such as the DWT and comes with a redundancy factor of 2.8 when wavelets are chosen at the finest scale and 7.2, otherwise [90].

### 3.1.3 Contourlet Transform

As seen in the previous subsection, the CVT tiles the 2-D frequency plane using the polar coordinate system. This makes its construction simple in the continuous domain but causes the implementation for discrete images - sampled on a rectangular grid - to be very challenging. This fact motivated the development of a further directional multiscale transform called Contourlet Transform (ConT) which, unlike the CVT, is defined directly in the discrete domain. It is worth emphasizing that, although the ConT and CVT have some similar properties and goals, the former is not a discretized version of the latter. Generally speaking, the ConT consists of a double filter bank structure where, first, a Laplacian pyramid (LP) [91] is used to

capture the point discontinuities before applying a directional filter bank (DFB) [92] which links the point discontinuities into linear structures. In what follows these two building blocks are introduced in more detail.

One of the earliest multiscale approaches in image processing is the pyramid representation. A classical image pyramid consists of a sequence of versions of an original image in which the resolution is gradually decreased by filtering and downsampling. The bottom (or zero) level $x^0$ of the pyramid is equal to the original image $I$. This image is low-pass filtered and downsampled to obtain the next level $x^1$. Further repetitions of this filtering/downsampling procedure generate the subsequent levels of the pyramid. This can be expressed by

$$x^{j+1}[m,n] = \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} w[k,l] x^j[2m-k, 2n-l], \tag{3.20}$$

where $w$ are the filter coefficients and $j$ is the current decomposition level. In general $w$ is separable and can be expressed such that $w[m,n] = h[m]h[n]$. If $w$ is chosen in a way that it resembles a Gaussian function, the resulting pyramid is referred to as the Gaussian pyramid.

By interpolating each image $x^{j+1}$ of the Gaussian pyramid and subtracting it from its predecessor $x^j$ we obtain the LP. The interpolation operation is defined as

$$\hat{x}^j[m,n] = 4 \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} w[k,l] x^{j+1}\left[\frac{m-k}{2}, \frac{n-l}{2}\right], \tag{3.21}$$

where only the terms for which $m-k$ and $n-l$ are even are included in the summation. Note that the image $\hat{x}^j$ can be interpreted as a prediction of $x^j$. Thus, each detail image $y^{j+1}$ of the LP corresponds to the error of approximation

$$y^{j+1}[m,n] = x^j[m,n] - \hat{x}^j[m,n], \quad 0 \leq j \leq J \tag{3.22}$$

with the exception of the top-level detail image $y^{J+1}$, which is defined as

$$y^{J+1}[m,n] = x^J[m,n], \tag{3.23}$$

where $x^J$ is the coarsest image in the Gaussian pyramid.

The original image $I = x^0$ can be recovered exactly by interpolating $y^{J+1}$ and adding it to $y^J$ to form $x^{J-1}$. This procedure is repeated until $x^0$ is reached. Note that the LP does not represent a critically-sampled transform and comes with a redundancy of approximately 33%. Furthermore we would like to point out that the LP accounts for the multiscale property of the ConT.

The multidirectionality of the ConT is achieved by a maximally decimated DFB.

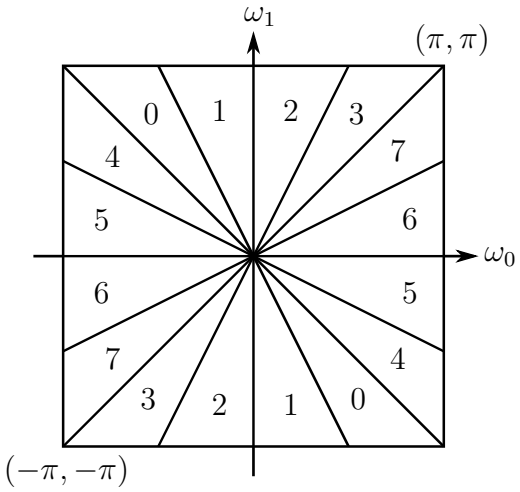



**Figure 3.4:** *Wedge-shaped frequency partition of the* 3*-level DFB.*

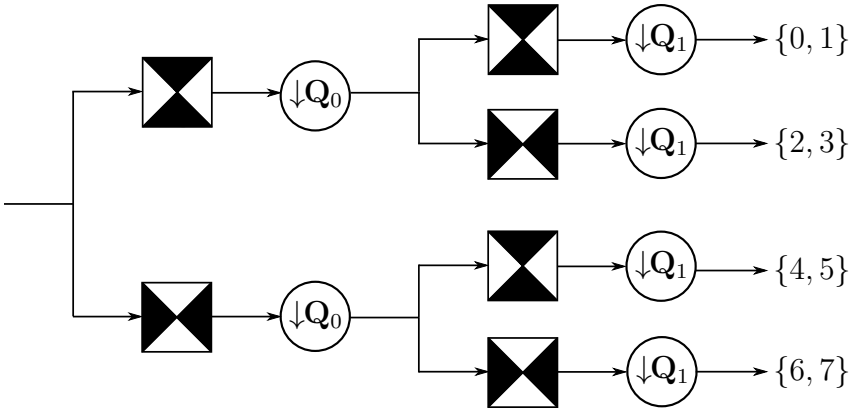



**Figure 3.5:** *The first two levels of the DFB. The black regions represent the ideal frequency support of the fan filters. The depicted set of numbers at the output of the DFB correspond to the directional subbands given in Fig. 3.4.*

It is intuitively constructed by combining fan filters [93] with resampling operations, leading to a $2^l$ wedge-shaped frequency partition as illustrated in Fig. 3.4 for $l = 3$.

To obtain a four directional frequency partitioning, the first two decomposition levels of the DFB are given in Fig. 3.5, where the sampling matrices $\mathbf{Q}_0$ and $\mathbf{Q}_1$ are defined as

$$\mathbf{Q}_0 = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}, \qquad \mathbf{Q}_1 = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}. \tag{3.24}$$

Note that $\mathbf{Q}_0\mathbf{Q}_1 = 2 \cdot \mathbf{I}_2$ with $\mathbf{I}_2$ donating the $2 \times 2$ identity matrix so that the overall sampling after two levels corresponds to a downsampling of two in each dimension. Using the noble identities of multirate systems [84] we can interchange the filters at the second level in Fig. 3.5 with the sampling matrix $\mathbf{Q}_0$. This change transforms the fan filter into a filter with checker-board frequency support (see second level of the nonsubsampled DFB in Fig. 3.8(b) for a schematic presentation of the idealized

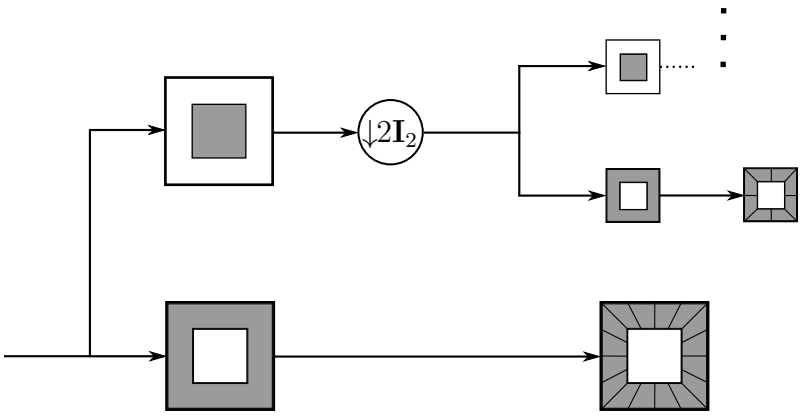



**Figure 3.6:** *ConT filter bank. First, a multiscale decomposition into octave bands by the LP is computed, and then a DFB is applied to each detail image.*

frequency support of these filters). The combination of the fan filters of the first level with the transformed filters of the second level results in four directional subbands as depicted at the output of the DFB in Fig. 3.5. From the third level onwards, to achieve finer frequency partition, the sampling matrices $\mathbf{Q}_0$ and $\mathbf{Q}_1$ are combined with the matrices

$$\begin{aligned} R_0 &= \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \qquad R_1 = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}, \\ R_2 &= \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}, \qquad R_3 = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}. \end{aligned} \tag{3.25}$$

After applying the noble identities, this leads to so-called parallelogram filters. In connection with the filters of the first and second level, this leads to eight directional subbands as shown in Fig. 3.4. The idealized frequency support of the parallelogram filters is illustrated at the third level of the nonsubsampled DFB in Fig. 3.8.

The ConT is obtained by simply concatenating the LP with the DFB, resulting in the overall double filter bank structure shown in Fig. 3.6. Due to the iterated directional filter bank approach, the ConT permits any number of $2^l$ directions at each scale. Moreover, like in case of the CVT if the number of directions is doubled at every other scale, the ConT satisfies the anisotropic scaling law, hence, it is able to efficiently approximate smooth objects having discontinuities along $C^2$ curves. We would like to point out that, due to the application of the LP, the ConT comes with a small redundancy of 33 percent. More detailed background information on the ConT can be found in [82].

### 3.1.4 Undecimated Wavelet Transform

While the decimated (bi)orthogonal wavelet transform is widely used in image compression algorithms such as JPEG-2000 [88], results are far from optimal for other applications like image fusion. This is mainly due to the downsampling in each decomposition step of the DWT which may cause a large number of artifacts when reconstructing an image after modification of its wavelet coefficients [94]. Thus, for applications such as image fusion, where redundancy is not a crucial factor, performance can be improved significantly by removing the decimation step in the DWT, leading to the non-orthogonal, translation-invariant Undecimated Wavelet Transform (UWT).

Like the DWT, the UWT is implemented using a filter bank which decomposes a discrete 1-D signal $x^0$ into a set $\mathcal{S} = \{y^1, \ldots, y^J, x^J\}$ in which $y^j$ represent the detail (or wavelet) coefficients at scale $j$ and $x^J$ are the approximation (or scaling) coefficients at the coarsest scale $J$. The passage from one resolution to the next one is obtained using the "à trous" algorithm [94, 95], where the analysis low-pass and analysis high-pass filter $h$ and $g$ are upsampled by $2^j$ when processing the $j^{\text{th}}$ scale and $j = 0, \ldots, J$. Thus, the UWT decomposition is defined as

$$x^{j+1}[n] = \sum_{l=-\infty}^{\infty} h[2^j l - n] x^j[l] = \left(\bar{h}^{(j)} * x^j\right)[n], \tag{3.26}$$

$$y^{j+1}[n] = \sum_{l=-\infty}^{\infty} g[2^j l - n] x^j[l] = \left(\bar{g}^{(j)} * x^j\right)[n], \tag{3.27}$$

where $\bar{h}[n] = h[-n]$ and $h^{(j)}[n] = h[\frac{n}{2^j}]$ if $\frac{n}{2^j}$ is an integer and 0, otherwise. The reconstruction at scale $j$ is obtained by

$$x^j[n] = \frac{1}{2}\left[\left(\tilde{h}^{(j)} * x^{j+1}\right)[n] + \left(\tilde{g}^{(j)} * y^{j+1}\right)[n]\right], \tag{3.28}$$

where $\tilde{h}$ and $\tilde{g}$ are the upsampled low-pass and high-pass synthesis filters, respectively.

The original signal can be recovered exactly from its UWT decomposition if the used analysis and synthesis filters satisfy the perfect reconstruction condition of eq. (3.8). This provides additional freedom during the filter selection process compared to the DWT where, in addition to the perfect reconstruction condition, the anti-aliasing condition of eq. (3.9) has to be satisfied as well.

The UWT can be extended to 2-D by filtering the rows and columns separately by $h$ and $g$ as given in eqs. (3.10) to (3.13), leading to three oriented detail images that isolate the horizontal, vertical and diagonal directions. The redundancy factor of an UWT $J$-level decomposition is $3J + 1$, since each detail image has the same

size than the original image.

Due to the fact that the filters do not need to be (bi)orthogonal, an alternative approach in multispectral image fusion (e.g. fusion of high-resolution panchromatic images with low-resolution multispectral images) is to define $g[n] = \delta[n] - h[n]$, where $\delta[n]$ represents an impulse at $n = 0$ [20, 22, 23]. In 2-D this yields $g[m, n] = \delta[m, n] - h[m, n]$, which suggests that the detail images can be obtained by taking the difference between two successive approximation images

$$y^{j+1}[m, n] = x^j[m, n] - x^{j+1}[m, n]. \tag{3.29}$$

Please note that, in this case, we only obtain one detail image for each scale and not three as in the general case. The reconstruction is obtained by co-addition of all detail images to the approximation image, that is

$$x^0[m, n] = x^J[m, n] + \sum_{j=1}^{J} y^j[m, n], \tag{3.30}$$

which implies that the synthesis filters are all-pass filters with $\tilde{h}[m, n] = \tilde{g}[m, n] = \delta[m, n]$ [94]. A common choice for the analysis, low-pass filter $h$ is a B-spline filter. In the literature this implementation of the UWT is known as Isotropic Undecimated Wavelet Transform [94] or Additive Wavelet Transform [22].

### 3.1.5 Dual-Tree Complex Wavelet Transform

Despite the success of classical wavelet methods, some limitations reduce their effectiveness in certain situations. For example the DWT and the UWT rely on a dictionary of roughly isotropic elements and contain only a small number of directions due to the standard tensor product construction in 2-D. Moreover, since wavelets are bandpass functions, their coefficients tend to oscillate around singularities which might complicate singularity extraction [96, 97].

A transform which attempts to circumvent the shortcomings of wavelet-based transforms is the Dual-Tree Complex Wavelet Transform (DTCWT). One way to understand the DTCWT is to note that the Fourier Transform does not suffer from many of the problems associated with wavelet transforms. The reason is that, unlike wavelet transforms, the Fourier Transform is based on complex-valued oscillating sinusoids

$$e^{j\omega n} = \cos(\omega n) + j \sin(\omega n) \tag{3.31}$$

which form a Hilbert transform pair (they are 90° out of phase with each other), thus constituting an analytic signal that is supported on only one-half of the frequency axis ($\omega > 0$). The DTCWT attempts to imitate the behavior of the Fourier Trans-

form by defining a complex-valued scaling and complex-valued wavelet function

$$\begin{array}{rcl} \phi_c(n) & = & \phi_r(n) + j\phi_i(n) \\ \psi_c(n) & = & \psi_r(n) + j\psi_i(n) \end{array}, \tag{3.32}$$

where $\phi_r$ and $\phi_i$ as well as $\psi_r$ and $\psi_i$ are (approximately) 90° out of phase with each other. By computing the inner product between the input signal $x^0$ and the complex-valued wavelet function $\psi_c$ the complex detail coefficients

$$y_c^j[n] = y_r^j[n] + jy_i^j[n] \tag{3.33}$$

with magnitude

$$\left|y_c^j[n]\right| = \sqrt{y_r^j[n]^2 + y_i^j[n]^2} \tag{3.34}$$

and phase

$$\angle y_c^j[n] = \arctan\left(\frac{y_i^j[n]}{y_r^j[n]}\right) \tag{3.35}$$

are obtained. Here, a large magnitude indicates the presence of a singularity while the phase indicates its position within the support of the wavelet [96]. Note, that the complex approximation coefficients are defined similarly.

The implementation of the DTCWT is done using two filter bank trees, each employing a real DWT, thus giving the transform its name. The first tree represents the real part of the transform while the second tree gives the imaginary part. The two trees use two different sets of filters, with each set satisfying the perfect reconstruction conditions of eqs. (3.8) and (3.9). In order for the two filter pairs to be analytic, the two analysis low-pass filters $h_0$ and $h_1$ should be a half-sample delay of the other

$$h_1[n] \approx h_0[n - 0.5], \tag{3.36}$$

which can only be satisfied approximately since $h_0[n]$ and $h_1[n]$ are defined solely for integers [96]. However, we can make the statement rigorous using the Fourier transform

$$H_1\left(e^{j\omega}\right) = e^{-j0.5\omega} H_0\left(e^{j\omega}\right). \tag{3.37}$$

The reconstruction of the original signal is obtained by simply inverting the two real DWTs, using the corresponding synthesis filters, and averaging the outputs of each tree.

It turns out that for the half-sample delay condition of eqs. (3.36) and (3.37) to be satisfied any perfect reconstruction filter bank can be used at the first decomposition stage. It is only necessary to translate the filter bank of one tree by one sample with respect to the filter bank of the other tree. As for all further stages so-called quarter

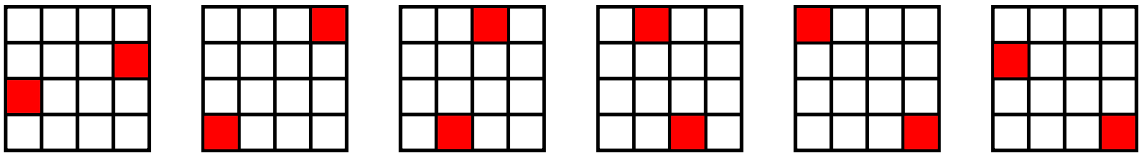


**Figure 3.7:** *Idealized support of the six oriented wavelets of the DTCWT in the 2-D frequency plane.*

sample shift (q-shift) orthogonal filter banks are used.

In addition, due to the (approximately) analytic nature of the DTCWT it is possible in 2-D to obtain complex wavelets which cover more distinct directions than the DWT. Fig. 3.7 illustrates this, showing the idealized support of each wavelet in the frequency domain. Note that, in 2-D, independent of the number of decomposition levels, the DTCWT is four times redundant.

Another attractive property of the DTCWT is its near shift-invariance which implies that a small shift of the input signal only results in a corresponding translation of the transform coefficients. This is in contrast to other transforms such as the DWT, CVT and ConT where transform coefficients may disappear arbitrarily under image translation. We see later in this work that shift-invariance (or translation-invariance) is a desirable property in image fusion applications. The interested reader can find more information on the DTCWT in [96] and [98].

### 3.1.6 Nonsubsampled Contourlet Transform

The Nonsubsampled Contourlet Transform (NSCT) represents the undecimated, shift-invariant counterpart of the ConT. Like the ConT, it can be conceptually divided into two parts: a) a nonsubsampled pyramid structure that ensures the multiscale property and b) a nonsubsampled DFB structure that gives directionality [99].

The pyramidal decomposition is obtained by removing the downsamplers and upsamplers of the Laplacian pyramid (LP) described in Section 3.1.3. This is in spirit similar to the Isotropic Undecimated Wavelet Transform where also a two-channel nonsubsampled filter bank is employed, yielding one detail image per stage. The filters at every subsequent stage are obtained by upsampling those of the previous stage according to the "à trous" algorithm. The idealized frequency support of the three-stage, nonsubsampled pyramid decomposition of a sample image $x^0$ is illustrated in Fig. 3.8(a), where the dark-gray areas correspond to the support of the deployed filters.

After decomposing the original image using the nonsubsampled LP, the resulting detail images are further processed using a nonsubsampled directional filter bank (DFB). In analogy to the nonsubsampled pyramid decomposition, it is constructed

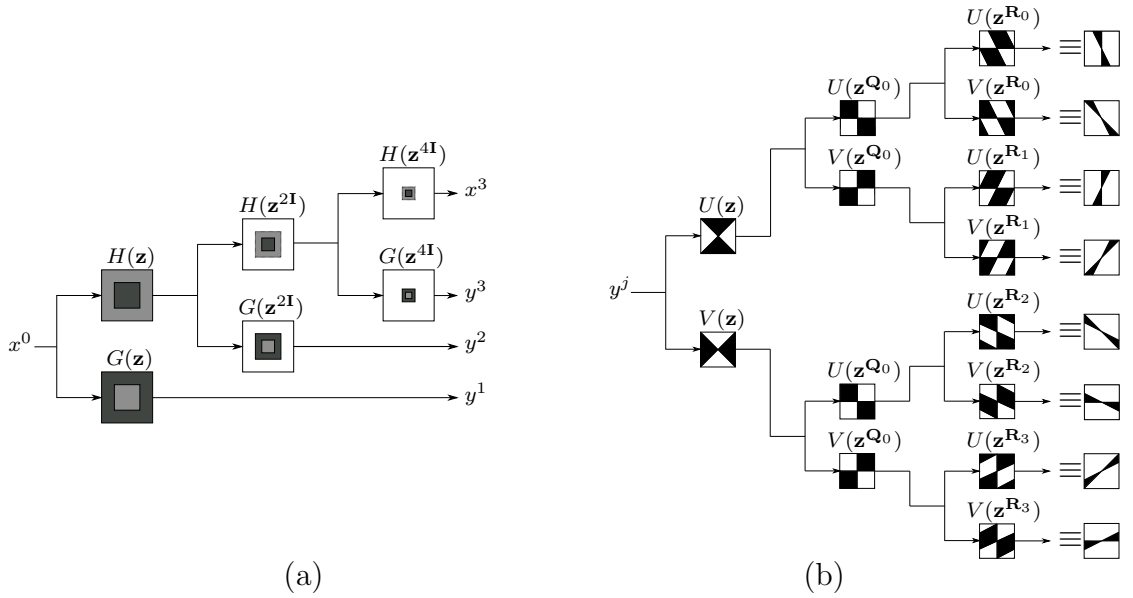



**Figure 3.8:** *Idealized frequency support of the two building blocks of the NSCT. (a) Nonsubsampled pyramid decomposition and (b) Nonsubsampled DFB.*

by eliminating the subsampling operations of the critically sampled DFB introduced in Section 3.1.3. A schematic diagram of the 3-level nonsubsampled DFB is illustrated in Fig. 3.8(b), showing the idealized frequency support of the resampled filters at each stage. Note that $U(z)$ and $V(z)$ correspond to the prototype fan filter pair in the $z$-transform domain that divides the 2-D frequency spectrum into horizontal and vertical directions, respectively. The resampling matrices $Q_0$, $R_0$, $R_1$, $R_2$ and $R_3$ are the same than those utilized in the critically sampled DFB (see eqs. (3.24) and (3.25)). The NSCT has a redundancy factor of $1 + \sum_{j=1}^{J} 2^{l_j}$, where $l_j$ denotes the number of levels of the nonsubsampled DFB at the $j^{\text{th}}$ scale. ∎

Before presenting the fusion results obtained for each multiscale transform introduced in this section, we first elaborate on the question how the "quality" of a fused image can be measured. This issue represents an inherent problem of image fusion since in most scenarios no ideal fusion result is available that may be used as ground truth.

## 3.2 Objective performance evaluation

The widespread use of image fusion has led to an increasing need for pertinent quality assessment tools in order to compare results obtained with different algorithms or to obtain an optimal setting of parameters for a given fusion algorithm. In general we are interested in measures that express the success of an image fusion technique in creating a composite image that retains as much salient information as possible from the source images while minimizing the number of artifacts that may incommode

human observers [7].

The most reliable and trusted methods of fusion assessment are subjective or perceptual image fusion evaluation trials in which an audience of potential users is employed to evaluate the fusion system under test. However, subjective tests are impractical in many cases due to heavy organizational and equipment requirements and strict test conditions that have to be obeyed. Objective fusion metrics that require no display equipment and are less time-consuming are therefore highly desirable. Of particular interest are fully automatic, non-reference fusion metrics which evaluate fusion without presuming knowledge of a ground truth as e.g. is needed for the root mean squared error evaluation [40, 45] of multifocus fusion. These metrics consider only the input images and the fused image to produce a single numerical score that indicates the success of the fusion process [100]. In the remainder of this section three of the most frequently used non-reference fusion metrics, namely, the performance measure proposed by Xydeas and Petrović $Q_{AB/F}$ [101], Piella $Q_P$ [102] as well as the Mutual Information (MI), first introduced by Qu et al. [103] in the context of image fusion, are discussed. Please note that these metrics are also employed in the remainder of this work to objectively assess the obtained fusion results. An exhaustive study on different objective metrics in the context of image fusion can be found in [104].

### 3.2.1 $Q_{AB/F}$

The $Q_{AB/F}$ fusion metric associates important visual information with the edge information present in an image. Thus, a fused image containing the edge information from all input images is considered to be the ideal fusion result.

Consider two input images $I_A$ and $I_B$, together with a fused image $I_F$. The $Q_{AB/F}$ fusion metric starts by applying a Sobel edge operator to the input image pair yielding the edge strength $g[m,n]$ and orientation information $\alpha[m,n]$ at each pixel position $m,n$. For the case of $I_A$ these values are defined as

$$g_{I_A}[m,n] = \sqrt{s^H_{I_A}[m,n]^2 + s^V_{I_A}[m,n]^2} \tag{3.38}$$

and

$$\alpha_{I_A}[m,n] = \tan^{-1}\left(\frac{s^V_{I_A}[m,n]}{s^H_{I_A}[m,n]}\right) \tag{3.39}$$

where $s^H_{I_A}[m,n]$ and $s^V_{I_A}[m,n]$ are the outputs of the horizontal and vertical Sobel operator, respectively.

Next, the relative edge strength and orientation values $G[m,n]$ and $A[m,n]$, respectively, between the input images and the fused image $I_F$ are calculated. For

$I_A$ this yields

$$G_{I_A I_F}[m,n] = \begin{cases} \frac{g_{I_F}[m,n]}{g_{I_A}[m,n]} & \text{if } g_{I_A}[m,n] > g_{I_F}[m,n] \\ \frac{g_{I_A}[m,n]}{g_{I_F}[m,n]} & \text{otherwise} \end{cases} \tag{3.40}$$

and

$$A_{I_A I_F}[m,n] = 1 - \frac{|\alpha_{I_A}[m,n] - \alpha_{I_F}[m,n]|}{\pi/2}. \tag{3.41}$$

These values are now used to derive the so-called edge strength and orientation preservation values $Q^g[m,n]$ and $Q^\alpha[m,n]$ which model the perceptual loss of information in the fused image $I_F$ in terms of how well the relative strength and orientation values $G[m,n]$ and $A[m,n]$ are represented in the fused image. For the input image $I_A$ these values are given by

$$Q^g_{I_A I_F}[m,n] = \frac{\Gamma_g}{1 + e^{\kappa_g \left(G_{I_A I_F}[m,n] - \sigma_g\right)}} \tag{3.42}$$

and

$$Q^\alpha_{I_A I_F}[m,n] = \frac{\Gamma_\alpha}{1 + e^{\kappa_\alpha \left(A_{I_A I_F}[m,n] - \sigma_\alpha\right)}}. \tag{3.43}$$

where the constants $\Gamma_g$, $\kappa_g$, $\sigma_g$ and $\Gamma_\alpha$, $\kappa_\alpha$, $\sigma_\alpha$ are used to determine the exact shape of the corresponding sigmoid functions. The overall edge information preservation value between the input image $I_A$ and the fused image $I_F$ is then defined as

$$Q_{I_A I_F}[m,n] = Q^g_{I_A I_F}[m,n] Q^\alpha_{I_A I_F}[m,n]. \tag{3.44}$$

Finally, the overall fusion metric $Q_{AB/F}$ is defined as

$$Q_{AB/F} = \frac{\sum_{m=1}^{M} \sum_{n=1}^{N} Q_{I_A I_F}[m,n] g_{I_A}[m,n] + Q_{I_B I_F}[m,n] g_{I_B}[m,n]}{\sum_{m=1}^{M} \sum_{n=1}^{N} g_{I_A}[m,n] + g_{I_B}[m,n]}, \tag{3.45}$$

where $g[m,n]$ is the result of applying a Sobel edge operator to the input image pair as given in eq. (3.38). Note that the resulting fusion score falls within a range of 0 and 1, with 0 representing total loss of edge information and 1 ideal fusion. In practice, however, this range is much narrower and small differences may indicate significant changes when perceptually analyzing the fused images.

### 3.2.2 Mutual Information

The Mutual Information (MI) fusion metric is a simple adaption of the mutual information concept of information theory. In the context of image fusion, the MI measure indicates how much information the composite image conveys about each of the source images. It is defined by simply adding the mutual information between the composite image $I_F$ and each of the input images $I_A$ and $I_B$ such that

$$\mathrm{MI} = I(I_A; I_F) + I(I_B; I_F), \tag{3.46}$$

where $I(I_k; I_F)$ is given by

$$I(I_k; I_F) = \sum_{u=1}^{L} \sum_{v=1}^{L} p_{I_k I_F}[u, v] \log_2 \frac{p_{I_k I_F}[u, v]}{p_{I_k}[u] p_{I_F}[v]}. \tag{3.47}$$

Here, $p_{I_k}$ and $p_{I_F}$ are the normalized gray level histograms of $I_k$ and $I_F$, respectively, $p_{I_k I_F}$ is the joint gray level histogram of $I_k$ and $I_F$, and $L$ is the number of bins (e.g. 256). Thus, the higher the mutual information between the fused image and the input images, the better the composite image resembles the ideal fusion result. Please note that in order to remove the dependency of the final score on the entropy of the input images, as well as to bound the final result to the interval $[0, 1]$, we additionally divided the metric by the sum of the individual entropies of the input images.

### 3.2.3 $Q_P$

Piella's performance metric $Q_P$ is based on the structural similarity (SSIM) index [105], introduced by Wang et al. and is given by the following expressions

$$Q_W(A, B, F) = \sum_{\mathbf{n} \in \mathcal{S}} c(\mathbf{n}) \left( \frac{s(A, \mathbf{n}) Q_0(A, F, \mathbf{n})}{s(A, \mathbf{n}) + s(B, \mathbf{n})} + \frac{s(B, \mathbf{n}) Q_0(B, F, \mathbf{n})}{s(A, \mathbf{n}) + s(B, \mathbf{n})} \right) \tag{3.48a}$$

$$Q_P(I_A, I_B, I_F) = Q_W(I_A, I_B, I_F) \cdot Q_W(\nabla I_A, \nabla I_B, \nabla I_F), \tag{3.48b}$$

with

$$c(\mathbf{n}) = \frac{\max\big(s(A, \mathbf{n}), s(B, \mathbf{n})\big)}{\sum_{\mathbf{m} \in \mathcal{S}} \max\big(s(A, \mathbf{m}), s(B, \mathbf{m})\big)}, \tag{3.49}$$

where $\{A, B, F\}$ represents a set of input images which either consists of the source images $I_A, I_B$ and the fused image $I_F$ or their corresponding gradient images $\nabla I_A$, $\nabla I_B$ and $\nabla I_F$, and $\mathcal{S}$ is the whole image.

In her approach, first two SSIM maps $Q_0$ are calculated, expressing the similarity between the first source image and the fused image as well as between the second

source image and the fused image. The SSIM index between the input image $I_A$ and the fused image $I_F$ at pixel position $\mathbf{n}$ is defined as

$$Q_0(I_A, I_F, \mathbf{n}) = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \cdot \frac{2\bar{x}\bar{y}}{\bar{x}^2 + \bar{y}^2} \cdot \frac{2\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}, \tag{3.50}$$

where $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ are two real-valued sequences, representing the gray-scale values of $I_A$ and $I_F$, respectively, within an arbitrary window centered at pixel position $\mathbf{n}$. $\bar{x}$, $\bar{y}$ and $\sigma_x^2$, $\sigma_y^2$ denote the mean and variance of the two sequences $x$ and $y$, respectively, and $\sigma_{xy}$ is the covariance of $x$ and $y$.

The two SSIM maps are afterwards refined by giving more weight to areas with a higher (perceptual) importance, according to some saliency measure $s$. In our implementation this saliency is expressed as the variance between the two input images $I_A$, $I_B$ and the fused image $I_F$, respectively, within a window of size $8 \times 8$.

Finally, $Q_W$ is calculated by pooling and averaging the weighted SSIM values of both maps, resulting in a score between -1 and 1, where values closer to 1 indicate a higher quality of the composite images. Furthermore, in order to take the importance of edge information into account, the above mentioned procedure is also applied to the corresponding gradient images $\nabla I_A$, $\nabla I_B$ and $\nabla I_F$. The final $Q_P$ fusion metric is calculated by multiplying the $Q_W$ score obtained from the original images with the $Q_W$ score coming from the gradient images.

### 3.2.4 Objective metric validation

The fusion metrics described above represent three different paradigms on how an ideally fused image should be assembled from an arbitrary set of source images. However, in order for these metrics to be truly applicable, their perceptual significance has to be established.

Such a validation was done in [100] where the author compared the objective fusion scores of the $Q_{AB/F}$, MI and $Q_P$ with the results of eight subjective trials including a total of 109 participants for 9 different multiscale fusion algorithms. As for the subjective evaluation trials, the participants were shown a series of image sets consisting of two inputs and two fused alternatives of these inputs. For each image set, the subjects were asked to express their individual preference for one or none of the fused images offered. In order to quantify the correlation between the objective and subjective evaluation results, two distinct correspondence measures were defined. The first measure, entitled correct ranking measure $CR$, evaluates the ability of an objective metric to predict the subjective preference for one of the two fused images offered for a particular input image pair. Thus, it expresses the proportion in which the subjective and objective ranking correspond. The second

| Objective metric | $Q_{AB/F}$ | MI | $Q_P$ |
|---|---|---|---|
| $r$ | 0.833 | 0.742 | 0.737 |
| $CR$ | 0.725 | 0.625 | 0.633 |

**Table 3.1:** *Subjective correspondence of the three objective fusion performance metrics $Q_{AB/F}$, MI and $Q_P$.*

measure $r$ is implemented in a similar fashion than the $CR$ measure but additionally takes into account the relative certainty of the subjective scores. For example, if all participants of the subjective trial unanimously voted for the same fusion result, a high correlation between the subjective and objective ranking is considered more crucial and is thus given more weight than in the case where the fused image received an evenly distributed number of votes.

Table 3.1 gives the subjective correspondence of the three fusion metrics $Q_{AB/F}$, MI and $Q_P$ in terms of the correct ranking $CR$ and subjective relevance $r$ scores, as presented in [100]. Note that both scores are bounded to the interval $[0, 1]$ with 1 indicating that the metric agrees in all cases with the subjective evaluation. From these results it is evident that all objective fusion scores correlate well with the results achieved in subjective trails. Thus, they can indeed be used to assess the suitability of different image fusion schemes. ■

Even though the perceptual effectiveness of all three employed fusion metrics could be established successfully, some open problems remain. For example, an ideal image fusion metric should not change with the content of the input images but rather evaluate the success of the fusion algorithm in creating an "ideal" fused image - a requirement which is satisfied by none of these metrics. Additionally, even though the fusion metrics considered in this work are bounded to the interval $[0, 1]$ ($Q_{AB/F}$ and MI) and $[-1, 1]$ ($Q_P$), respectively, the distribution within this range is not clear. In other words, given two metric values 0.99 and 0.96, for instance, we do not know how significant the difference 0.03 is.

Thus, based on these two examples, we can conclude that there still exists a pertinent need for new fusion metrics which do not suffer from any of these shortcomings and consequently allow for stronger assertions regarding the overall quality of the fused image.

## 3.3 Results

In this section, we compare the performance of different multiscale transforms using three fusion scenarios. The first scenario considers the fusion of images with different

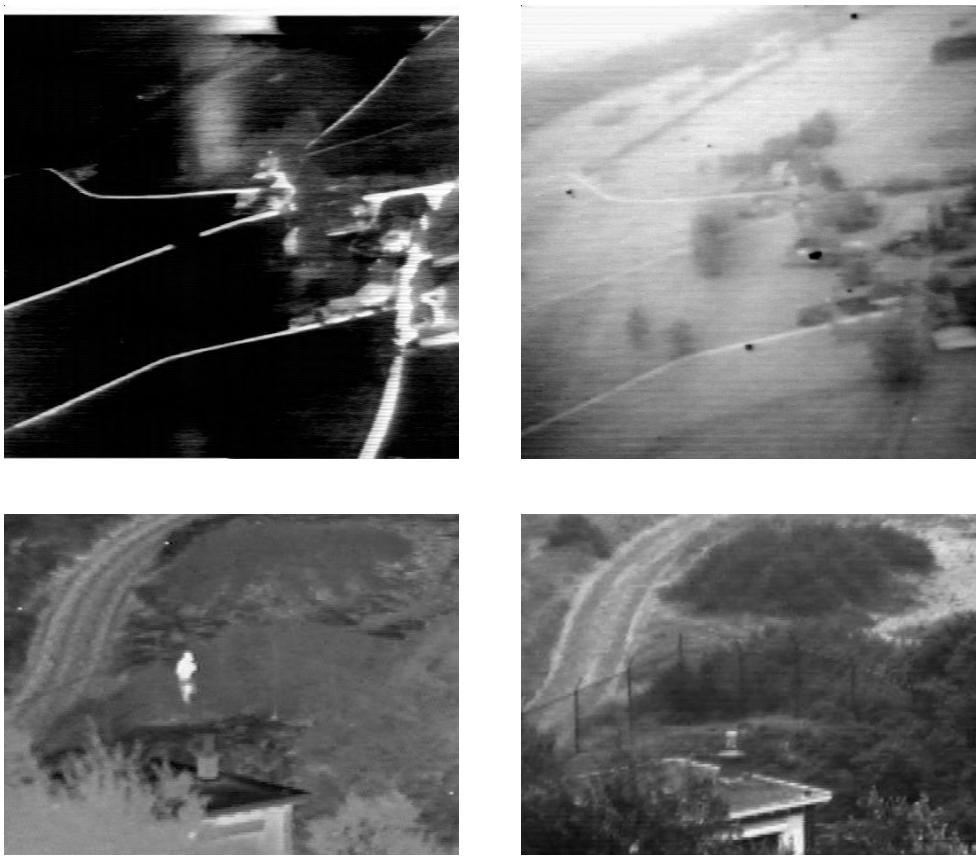


**Figure 3.9:** *Two IR-visible image pairs used for evaluation purposes. Left column consists of IR images, whereas the right column shows the corresponding visible images. Source images kindly provided by Dr. Oliver Rockinger and TNO, The Netherlands, respectively.*

focus points whereas the second and third scenario deal with the fusion of IR-visible and medical image pairs, respectively. All utilized source images, arranged in correspondence to their underlying fusion scenario, are illustrated in Figs. 3.9 to 3.11.

In all of our simulations the decomposed detail images $y_A^j$ and $y_B^j$ are fused using a simple "choose max" fusion rule. As discussed in Section 2.3, by this rule the coefficient yielding the highest energy is directly transferred to the fused decomposed representation. Hence, the fused, detail images $y_F^j$ are defined as

$$y_F^j[m,n,p] = \begin{cases} y_A^j[m,n,p] & \text{if } \left|y_A^j[m,n,p]\right| > \left|y_B^j[m,n,p]\right| \\ y_B^j[m,n,p] & \text{otherwise} \end{cases}, \tag{3.51}$$

where $m, n$ represent the spatial location in a given orientation band $p$ at decomposition level $j$. This choice is motivated by the fact that salient features result in large magnitude coefficients, and thus can be effectively captured using this fusion scheme. The low-pass approximation images $x_A^J$ and $x_B^J$ are treated differently since high magnitudes in the approximation images do not necessarily correspond to im-

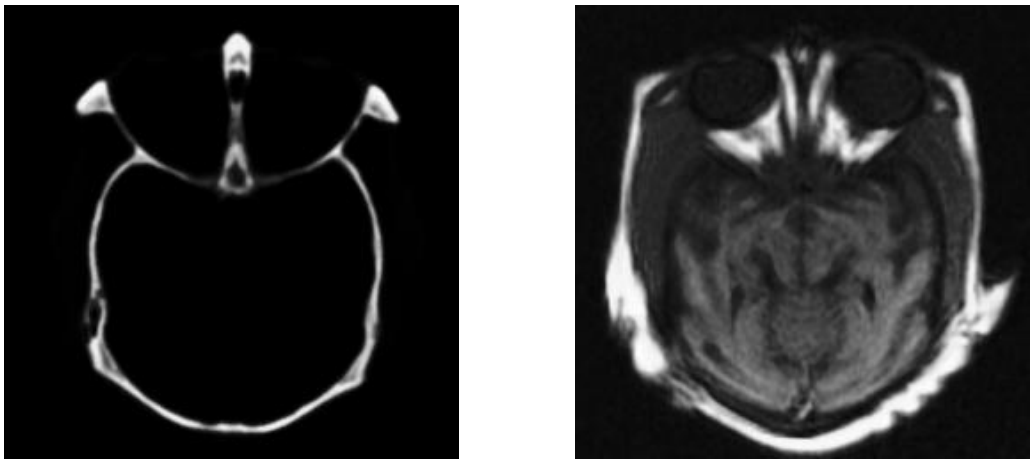


**Figure 3.10:** *Medical image pair used for evaluation purposes.*

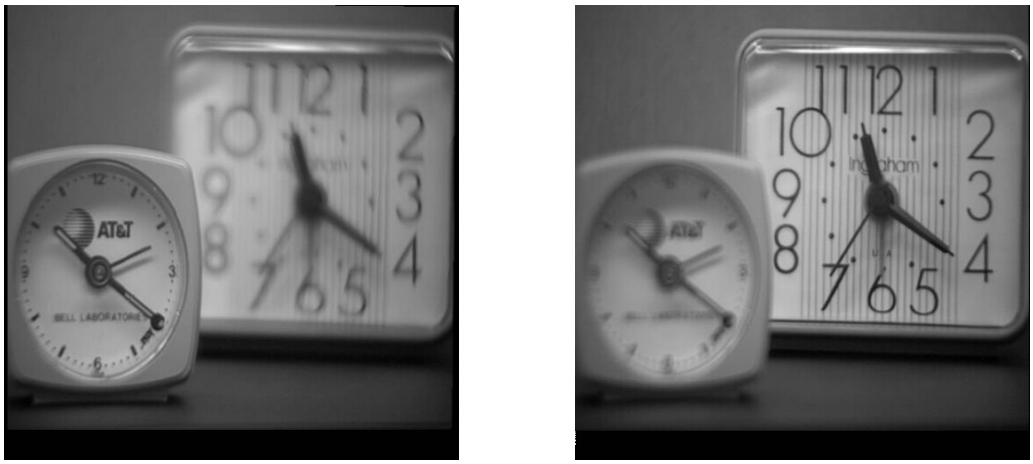


**Figure 3.11:** *Multifocus image pair used for evaluation purposes.*

portant features within the source images. Thus, in our experiments, the composite approximation image $x_F^J$ is obtained by a simple averaging operation,

$$x_F^J[m,n] = \frac{x_A^J[m,n] + x_B^J[m,n]}{2}. \tag{3.52}$$

In the literature more sophisticated fusion rules can be found. However, since the focus of this chapter is on assessing the suitability of different multiscale transforms for image fusion rather than on the used fusion rule, the "choose max" rule in combination with an averaging of the approximation images suffices for our purposes.

The multiscale transforms investigated in this section include all previously introduced decompositions, namely, the DWT, CVT, ConT as well as the UWT, DTCWT and NSCT using various filter bank settings and number of directions (in case of the CVT, ConT and NSCT). All simulations were conducted in Matlab®. In all cases the number of tested decomposition levels varies from two to five. As for the objective evaluation of the obtained fusion results the three objective metrics of Section 3.2 are utilized. In the remainder of this section we first individually as-

| Fusion Scenario | Filter Bank | Levels | $Q_{AB/F}$ | MI | $Q_P$ |
|---|---|---|---|---|---|
| Infrared-visible | *bior2.2* | 5 | **0.5268** | 0.1155 | **0.7059** |
| | *bior2.2* | 2 | 0.4757 | **0.1372** | 0.6013 |
| Medical | *db1* | 5 | **0.6625** | 0.3461 | **0.5912** |
| | *bior2.2* | 2 | 0.5262 | **0.4053** | 0.4926 |
| Multifocus | *sym8* | 5 | **0.6368** | 0.4542 | 0.8663 |
| | *bior2.2* | 2 | 0.6000 | **0.4710** | 0.7988 |
| | *db4* | 5 | 0.6329 | 0.4505 | **0.8676** |

**Table 3.2:** *Summary of the best fusion results for the DWT.*

sess the fusion performance of each multiscale decomposition, before giving a global comparison of the achieved results.

### 3.3.1 Discrete Wavelet Transform

In order to assess the fusion performance of the DWT we consider four different wavelet families[1] in our simulations: Daubechies (dB$N$, $N = 1, 2, \ldots, 10$), Symlets (sym$N$, $N = 2, 3, \ldots, 10$), Coiflets (coif$N$, $N = 1, 2, \ldots, 5$) and Biorthogonal (bior$\{M.N\}, M.N = 1.3, 1.5, 2.2, 2.4, 2.6, 2.8, 3.1, 3.3, 3.5, 3.7, 3.9, 4.4, 5.5, 6.8$). Additionally, for each wavelet basis, the number of decomposition levels is varied from two to five.

The best results for each fusion metric, highlighted in bold, are presented in Table 3.2. Please note that in case of the IR-visible fusion scenario, all reported values correspond to the average values of the two image pairs depicted in Fig. 3.9. At first glance it can be noted that a good performance is achieved using Biorthogonal and Daubechies filters for all three fusion scenarios. As for the fusion of IR-visible and medical image pairs the best results are obtained by using the 'bior2.2' and 'db1' (or Haar) filter bank, respectively, which exhibit short support sizes. In general it can be deduced that for multisensor images better fusion results are achieved using short filters.

In addition, it can be noted that a high number of decomposition levels seems to provide better fusion scores than a low number. However, it would be presumptuous to deduce that a higher decomposition depth automatically leads to better results. In fact, it may produce low-resolution bands where neighboring features overlap. This gives rise to discontinuities in the composite multiscale representation and may introduce distortions such as blocking or 'ringing' artifacts in the final fused image. Furthermore, the ideal decomposition depth is also related to the support of

[1]In the course of this work filters are referred to by their respective names within the Matlab® Wavelet Toolbox™. More information can be found at http://www.mathworks.com/products/wavelet/.

the employed filter bank and the size of the relevant objects in the source images, thus making it highly application dependent. However, as a rule of thumb, it seems that good performance is achieved for four to five decomposition levels. Please note that the MI fusion metric associates the best fusion performance to the use of only two decomposition levels, independent of the underlying fusion scenario. This is so because, as reported in various studies [2, 102, 103, 106], the MI fusion metric constantly assigns the best ranking to the averaging fusion rule which, in our case, is used to fuse the approximation images. Thus, since a small decomposition depth implies that more information is fused using the averaging method, the MI fusion metric tends to favor smaller decomposition numbers. However, as long as the performance comparison is carried out at the same number of decomposition levels, MI has been shown to be a good indicator of the quality of multiscale image fusion [103].

### 3.3.2 Curvelet Transform

The CVT is a member of the new, emerging family of directional multiscale transforms which can be implemented using any multiple of 4 directional decompositions per scale. Please note that in this work we utilized a ready-made Matlab® implementation of the CVT which is available under `http://http://www.curvelet.org` in its latest version 2.1.2.

In order to test the influence of the number of directional decompositions on the final fusion result, we conduct experiments with a varying number of orientations. More specifically, we consider $4 \cdot N, N = 2, 3, \ldots, 8$ numbers of directional decompositions at the $2^{\text{nd}}$ coarsest scale. As for the remaining, finer scales, we derive the number of orientations from the anisotropic scaling law given in eq. (3.18) which implies that we have to double the number of directions at every other scale. For example, for 6 frequency scales (including the coarsest non-directional scale) and 8 directional decompositions at the $2^{\text{nd}}$ coarsest scale, we obtain the following set of orientations, assorted from coarsest to finest scale: $\{1, 8, 16, 16, 32, 32\}$. Please note that in this context the use of 6 frequency scales results in the same number of frequency partitions than in the case of filter bank-based transforms with 5 decomposition levels.

As mentioned in Section 3.1.2, the design of appropriate digital curvelets at the finest scale is not straightforward because of boundary/periodicity issues. For this purpose we compare two constructions of the CVT, one employing non-directional wavelets and the other one using curvelets at the coarsest scale. Furthermore, we performed tests for three to six frequency scales. This corresponds to the same decomposition depths used to assess transforms implemented via filter banks.

| Fusion Scenario | Directions | $Q_{AB/F}$ | MI | $Q_P$ |
|---|---|---|---|---|
| Infrared-visible | $\{1, 8, 16, 16, 32, 32\}$ | **0.5302** | 0.1158 | 0.7178 |
| | $\{1, 8, 16\}$ | 0.4671 | **0.1362** | 0.5977 |
| | $\{1, 8, 16, 16, 32\}$ | 0.5234 | 0.1183 | **0.7229** |
| Medical | $\{1, 8, 16, 16, 1\}$ | **0.6260** | 0.2250 | 0.5807 |
| | $\{1, 8, 16\}$ | 0.5091 | **0.3998** | 0.4678 |
| | $\{1, 8, 16, 16, 32\}$ | 0.6257 | 0.2251 | **0.5808** |
| Multifocus | $\{1, 8, 16, 16, 32, 32\}$ | **0.6649** | **0.4785** | **0.8831** |

**Table 3.3:** *Summary of the best fusion results for the CVT.*

Table 3.3 summarizes the best fusion results for the CVT. Note that the set of directions ($2^{\text{nd}}$ column) also provides information about the number of frequency scales and the utilized transform at the finest scale. For instance, the set of directions $\{1, 8, 16, 16, 1\}$ implies that five frequency scales with wavelets at the finest scale and 8 directional decompositions at the $2^{\text{nd}}$ coarsest scale are used. Similar to the DWT, we notice that in all fusion scenarios a higher number of frequency scales results in a better $Q_{AB/F}$ and $Q_P$ fusion score. For the IR-visible and medical fusion scenarios, the MI metric yields the best results for 3 frequency scales whereas for multifocus image fusion the best MI results are achieved using 6 frequency partitions. The last result is in strong contrast to the previous affirmation that the MI fusion metric always favors the lowest decomposition depth. In general, it can be observed that the performance of multifocus image fusion does not depend as strongly on the used number of frequency partitions than in the other tested fusion scenarios. We believe that this is due to the nature of multifocus image pairs which only differ in their high frequency content but are identical otherwise. More specifically, it seems that the differing information can already be captured successfully at coarser frequency scales, hence rendering finer frequency partitions less significant. As for the number of orientations per scale, the use of 8 directional decompositions at the $2^{\text{nd}}$ coarsest scale results in the best overall fusion performance. Moreover, it can be observed that Curvelets at the finest scale produce slightly better results than the use of wavelets. This shows that for image fusion applications, the unwanted spilling of directional information to other unrelated angles due to aliasing is not influencing the final result strongly enough to cancel out the advantages of directionality.

### 3.3.3 Contourlet Transform

As shown in Section 3.1.3, the implementation of the ConT is based on a double filter bank structure where the source images are first decomposed using the Laplacian Pyramid (LP) before applying a directional filter bank (DFB). This results in a

| Fusion Scenario | Pyramid Filter | DFB Filter | Directions | $Q_{AB/F}$ | MI | $Q_P$ |
|---|---|---|---|---|---|---|
| Infrared-visible | *5/3* | *9/7* | $\{1, 2, 2, 4, 4\}$ | **0.5113** | 0.1138 | 0.6914 |
| | *5/3* | *pkv12* | $\{1, 2\}$ | 0.4705 | **0.1355** | 0.6036 |
| | *5/3* | *pkv6* | $\{1, 2, 2, 4, 4\}$ | 0.5100 | 0.1134 | **0.6928** |
| Medical | *5/3* | *9/7* | $\{2, 4, 4, 8\}$ | **0.6196** | 0.2450 | 0.5832 |
| | *5/3* | *5/3* | $\{2, 4\}$ | 0.5361 | **0.4014** | 0.5029 |
| | *5/3* | *5/3* | $\{2, 4, 4, 8\}$ | 0.6163 | 0.2497 | **0.5852** |
| Multi-focus | *5/3* | *9/7* | $\{1, 2, 2, 4\}$ | **0.6350** | 0.4467 | 0.8697 |
| | *5/3* | *9/7* | $\{2, 4\}$ | 0.5991 | **0.4719** | 0.7900 |
| | *9/7* | *pkv12* | $\{1, 2, 2, 4, 4\}$ | 0.6257 | 0.4408 | **0.8740** |

**Table 3.4:** *Summary of the best fusion results for the ConT.*

multiscale and multidirectional decomposition similar to the CVT.

In order to assess the performance of the ConT, we utilize four different filters for the pyramid decomposition in conjunction with six (bi)orthogonal DFB prototype filters. This results in a total of 24 tested filter bank combinations. As for the LP, we implement the CDF 5/3 and CDF 9/7 filter banks, the original LP filter bank introduced by Burt in [91] and the 12-tap biorthogonal FIR filter proposed in [107]. Please note that in our simulations we refer to these filters as '5/3', '9/7', 'Burt' and 'pkv12', respectively. In case of the DFB, we employ the '5/3', '9/7' and 'pkv12' filter banks which are also used during the multiscale LP decomposition as well as the orthogonal Haar filter bank and the 6-tap and 8-tap FIR filter banks of [107]. In our experiments we address the three latter filter banks as 'Haar', 'pkv6' and 'pkv8 ', respectively. Note that the corresponding fan filters can be obtained by modulation of the six (bi)orthogonal prototype DFB filters.

The ConT allows for the implementation of any $2^l$ directions at each scale. Thus, in order to evaluate the influence of the number of directional decompositions, we perform experiments with $2^l, l = 0, 1, 2, 3$ directions at the coarsest decomposition stage. Like in the case of the CVT, the directions for the remaining decomposition levels are derived from the anisotropic scaling law, given in eq. (3.18). The number of tested multiscale decompositions ranges from two to five.

For each metric and fusion scenario, the best results are listed in Table 3.4, where the number of decomposition levels corresponds to the cardinality of the depicted set of directional decompositions. It can be seen that independent of the underlying fusion scenario and fusion metric, the best fusion performance is achieved by employing the CDF 5/3 filter bank during pyramid decomposition. As for the DFB, it is more difficult to give an explicit recommendation since several filter banks result in (almost) similar results. The only outlier is the Haar filter bank

| Fusion Scenario | Filter Bank | Levels | $Q_{AB/F}$ | MI | $Q_P$ |
|---|---|---|---|---|---|
| Infrared-visible | *db1* | 5 | **0.5716** | 0.1230 | 0.7160 |
| | *db1* | 2 | 0.5369 | **0.1422** | 0.6576 |
| | *db1* | 4 | 0.5704 | 0.1272 | **0.7234** |
| Medical | *db1* | 5 | **0.7302** | 0.2829 | **0.6559** |
| | *bior2.2* | 2 | 0.5910 | **0.4389** | 0.5183 |
| Multifocus | *db1* | 5 | **0.6806** | 0.4651 | 0.8787 |
| | *db1* | 2 | 0.6731 | **0.4864** | 0.8345 |
| | *db1* | 4 | 0.6786 | 0.4813 | **0.8804** |

**Table 3.5:** *Summary of the best fusion results for the UWT.*

which shows the worst fusion performance in all cases and can therefore be safely discarded. In addition, the best results are obtained by applying a comparably small number of directional decompositions. In fact in almost all cases the fusion scores deteriorate for an increase in the number of orientations. This phenomenon could also be observed in the case of the CVT. Finally, the highest scores for the $Q_{AB/F}$ and $Q_P$ are again obtained by employing four to five decomposition levels, whereas the MI yields the best performance for two decomposition levels.

The implementation of the ConT used in this work can be obtained at Matlab® Central `http://www.mathworks.com/matlabcentral/fileexchange/8837`.

### 3.3.4 Undecimated Wavelet Transform

To assess the performance of the UWT, we consider the same wavelet families and number of decomposition levels as in the case of the DWT.

Table 3.5 gives the best fusion results for the UWT. It can be seen that independent of the underlying fusion scenario and fusion metric, best results are achieved for the 'db1' (or Haar) filter bank which exhibits the shortest support size among all tested filters. In general, it can be deduced from the obtained results that the UWT fusion performance improves with decreasing filter support length. This applies especially to multisensor image fusion where vast differences between the source images can be observed. As previously, the fusion metrics $Q_{AB/F}$ and $Q_P$ produce the highest scores for four and five decomposition levels, respectively. In contrast, two decomposition levels result in the best fusion scores for the MI measure.

### 3.3.5 Dual-Tree Complex Wavelet Transform

The implementation of the DTCWT is done using a dual-tree filter bank where each tree employs a real DWT. In order for the two filter bank trees to form an approximately analytic decomposition, the half-sample delay condition of eq. (3.36)

| Fusion Scenario | Filter Banks | | Levels | $Q_{AB/F}$ | MI | $Q_P$ |
|---|---|---|---|---|---|---|
| | 1st stage | Other | | | | |
| Infrared-visible | *near_sym_a* | *qshift10-6* | 5 | **0.5671** | 0.1234 | 0.7307 |
| | *5/3* | *qshift10-6* | 2 | 0.4974 | **0.1411** | 0.6148 |
| | *near_sym_a* | *qshift18* | 5 | 0.5650 | 0.1219 | **0.7319** |
| Medical | *5/3* | *qshift10-6* | 4 | **0.6611** | 0.2619 | **0.6082** |
| | *9/7* | *qshift10-6* | 2 | 0.5414 | **0.4279** | 0.4891 |
| Multi-Focus | *5/3* | *qshift10-6* | 5 | **0.6718** | 0.4867 | **0.8865** |
| | *5/3* | *qshift10-10* | 5 | 0.6715 | **0.4875** | 0.8863 |

**Table 3.6:** *Summary of the best fusion results for the DTCWT.*

needs to be satisfied, resulting in so-called quarter sample shift (q-shift) orthogonal filter banks. In the course of our experiments we compare the performance of five such filter banks. More specifically, we utilize three q-shift filters with 18-, 16- and 14-taps, respectively, as well as two 10-tap q-shift filter banks. Note that the difference between the two 10-tap q-shift filters is that the first exhibits 10 non-zero taps whereas the second comes with only 6 non-zero taps. We refer to the five q-shift filters (in order of their appearance) as 'qshift18', 'qshift16', 'qshift14', 'qshift10-10' and 'qshift10-6', respectively. A comprehensive guide on the design of q-shift filter banks, including a detailed list of all filter bank coefficients, can be found in [108]. As discussed in Section 3.1.5, for eq. (3.36) to be satisfied at the first decomposition level it is sufficient to merely translate one set of filters by one sample with respect to the other one. Consequently, any perfect reconstruction filter bank can be used. In our simulations the CDF 5/3 ('5/3') and 9/7 ('9/7') filter banks as well as the near-symmetric 5/7 ('near_sym_a') and 13/19-tap ('near_sym_b') filters of [109] are employed at the first decomposition level. Note that we tested for all possible filter bank combinations, resulting in a total of 20 analyzed filter bank settings.

The best results for each fusion metric using two to five decomposition levels are depicted in Table 3.6. We can deduce from the depicted results that the best IR-visible fusion performance is achieved for the 'near_sym_a' filter bank in combination with the 'qshift10-6' filter bank. For medical and multifocus image fusion the best scores are obtained by employing the '5/3' filter bank at the first decomposition level and the 'qshift10-6' filter bank at all remaining stages. As in our previous experiments, four to five decomposition levels result in the best fusion results for the $Q_{AB/F}$ and $Q_P$ metrics. For the IR-visible and medical fusion scenarios, the MI metric yields the best results for 2 decomposition levels whereas for multifocus image fusion the best MI results are achieved using a decomposition depth of five.

As discussed in Section 3.3.2, this is due to the fact that fewer decomposition levels are needed to capture the differing information between the source images. In other words, a high decomposition depth is prone to produce detail images at coarser scales which are virtually identical.

A Matlab® implementation of the DTCWT can be obtained on inquiry from the main authors website `http://www-sigproc.eng.cam.ac.uk/~ngk`.

### 3.3.6 Nonsubsampled Contourlet Transform

The last decomposition under investigation is the NSCT. It is implemented using a concatenation of a nonsubsampled pyramid structure with a nonsubsampled directional filter bank (DFB) and represents the undecimated, shift-invariant counterpart of the Contourlet Transform. The implementation used in this work is available for download at `http://www.mathworks.com/matlabcentral/fileexchange/10049`.

In order to identify the best filter bank setup we tested for 40 different filter bank combinations. More specifically, we assessed the performance of four different pyramid filters, namely, the CDF 9/7 ('9/7') filter bank and the three maximally flat pyramid filters given in [99] (in our simulations referred to as 'maxflat1', 'maxflat2' and 'maxflat3', respectively). As for the fan filters used in the DFB construction, 10 different prototype filters were used. These are the orthogonal Haar filter, the CDF 9/7 filter bank, the 3/5-tap linear phase filter ('vk') given on page 143 of [86], the 19-tap 2-D non-separable diamond-shaped filter bank ('lax') of [110], the 9-tap filter ('sk') proposed in [111], the 12-tap FIR filter ('pkv12') of [107] and the diamond-shaped maximally flat filters of order 4, 5, 6 and 7 ('dmaxflat4', 'dmaxflat5', 'dmaxflat6' and 'dmaxflat7', respectively) described in [99]. Like in the case of the ConT, we performed tests with $2^l, l = 0, 1, 2, 3$ directions at the coarsest frequency scale and derived the directions for the remaining decomposition levels from the anisotropic scaling law of eq. (3.18). Furthermore, the number of tested multiscale decompositions ranged from two to five.

Table 3.7 lists the best NSCT fusion scores for each fusion metric and scenario. Please note that the cardinality of the set of directional decompositions corresponds to the utilized overall decomposition depth. By looking at the obtained IR-visible and medical fusion results it can be noted that the 'maxflat2' pyramid filter shows the best performance among all four tested filter banks. A decision for one of the pyramid filter banks in the multifocus fusion scenario seems to be more difficult since each fusion metric favors a different filter bank. However, by investigating the individual results more thoroughly it seems that the '9/7' pyramid filter produces slightly better results than the other competing filter banks. The '9/7' and the 'vk' filter banks appear to be the best choices during the directional decomposition

| Fusion Scenario | Pyramid Filter | DFB Filter | Directions | $Q_{AB/F}$ | MI | $Q_P$ |
|---|---|---|---|---|---|---|
| Infrared-visible | *maxflat2* | *9/7* | $\{2, 4, 4, 8, 8\}$ | **0.5735** | 0.1258 | 0.7296 |
| | *maxflat2* | *vk* | $\{8, 16\}$ | 0.5133 | **0.1425** | 0.6350 |
| | *maxflat1* | *9/7* | $\{4, 8, 8, 16, 16\}$ | 0.5698 | 0.1250 | **0.7308** |
| Medical | *maxflat2* | *vk* | $\{2, 4, 4, 8, 8\}$ | **0.7156** | 0.2860 | 0.6455 |
| | *maxflat2* | *sk* | $\{8, 16\}$ | 0.5941 | **0.4493** | 0.5167 |
| | *maxflat2* | *dmaxflat7* | $\{4, 8, 8, 16, 16\}$ | 0.7136 | 0.2862 | **0.6479** |
| Multi-focus | *9/7* | *vk* | $\{1, 2, 2, 4, 4\}$ | **0.6775** | 0.4787 | 0.8838 |
| | *maxflat3* | *vk* | $\{8, 16, 16, 32\}$ | 0.6747 | **0.4887** | 0.8812 |
| | *maxflat1* | *9/7* | $\{8, 16, 16, 32, 32\}$ | 0.6742 | 0.4817 | **0.8848** |

**Table 3.7:** *Summary of the best fusion results for the NSCT.*

of IR-visible and multifocus images. In case of medical image fusion, the decision proves again to be difficult. Nevertheless, in general the 'dmaxflat7' filter bank seems to perform better than the other contestants.

Furthermore, from the obtained results we see that two to eight directional decompositions at the coarsest scale produce good results for all fusion metrics. Note that this is substantially more than in case of the ConT, where the best results are obtained using at most one DFB stage. Regardless of the underlying fusion scenario, four to five multiscale decomposition levels produce the best performance for the $Q_{AB/F}$ and $Q_P$ measure. In case of the MI fusion metric, two decomposition levels yield the best results for IR-visible and medical image fusion whereas a decomposition depth of four shows the best performance in the multifocus scenario.

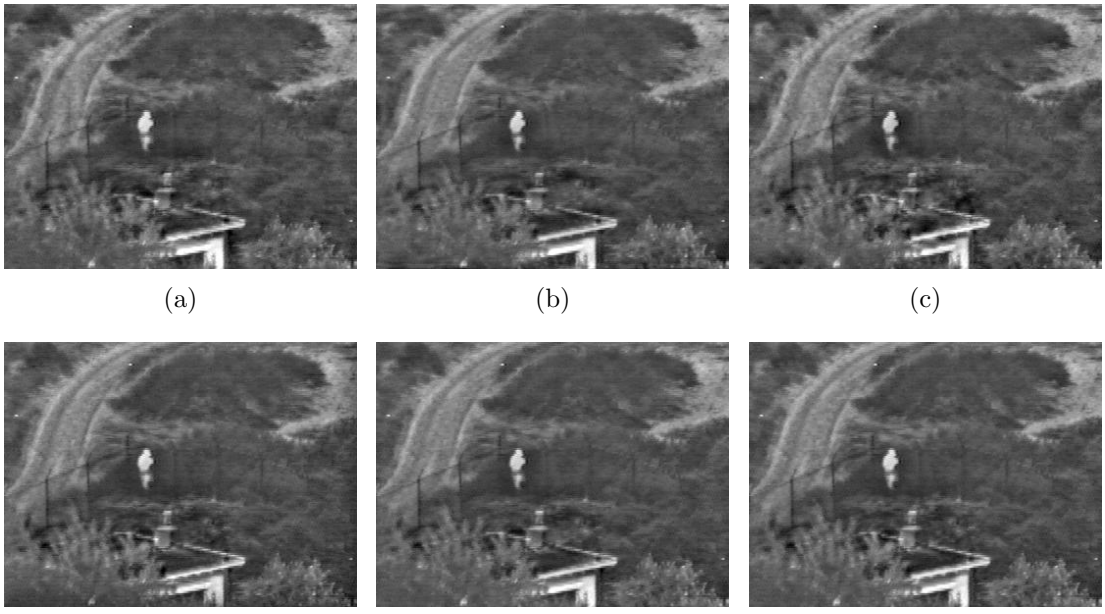
### 3.3.7 Global Comparison

In this section, the fusion results of all investigated multiscale transforms, divided into their underlying fusion scenarios, are compared and analyzed. For this purpose, Tables 3.8, 3.9 and 3.10 list the best global results for all IR-visible, medical and multifocus image pairs, respectively, obtained by applying the DWT, CVT, ConT, UWT, DTCWT and NSCT.

By analyzing the fusion results of Table 3.8 for the IR-visible fusion scenario, we observe that the best fusion scores are obtained for the NSCT followed by the DTCWT and the UWT. We attribute this fact mainly to the redundant, shift-invariant nature of these transforms. In general, it can be deduced that these features appear to be desirable properties in multiscale image fusion applications. This was also recognized in other studies such as [2], [5] and [23]. However, if one has to resort to shift-variant transforms with limited or no redundancy, e.g. for reasons of limited storage capacity, the CVT seems to be the best choice among them.

| Transform | Filter Bank(s) | | Levels/Directions | $Q_{AB/F}$ | MI | $Q_P$ |
|---|---|---|---|---|---|---|
| DWT | *bior2.2* | | 5 | 0.5268 | 0.1155 | 0.7059 |
| CVT | | | $\{1, 8, 16, 16, 32, 32\}$ | 0.5302 | 0.1158 | 0.7178 |
| ConT | *5/3* | *9/7* | $\{1, 2, 2, 4, 4\}$ | 0.5113 | 0.1138 | 0.6914 |
| UWT | *db1* | | 5 | 0.5716 | 0.1230 | 0.7160 |
| DTCWT | *near_sym_a* | *qshift10-6* | 5 | 0.5671 | 0.1234 | **0.7307** |
| NSCT | *maxflat2* | *9/7* | $\{2, 4, 4, 8, 8\}$ | **0.5735** | **0.1258** | 0.7296 |

**Table 3.8:** *Global comparison of the best results for the IR-visible image fusion scenario.*



(a) (b) (c)

(d) (e) (f)

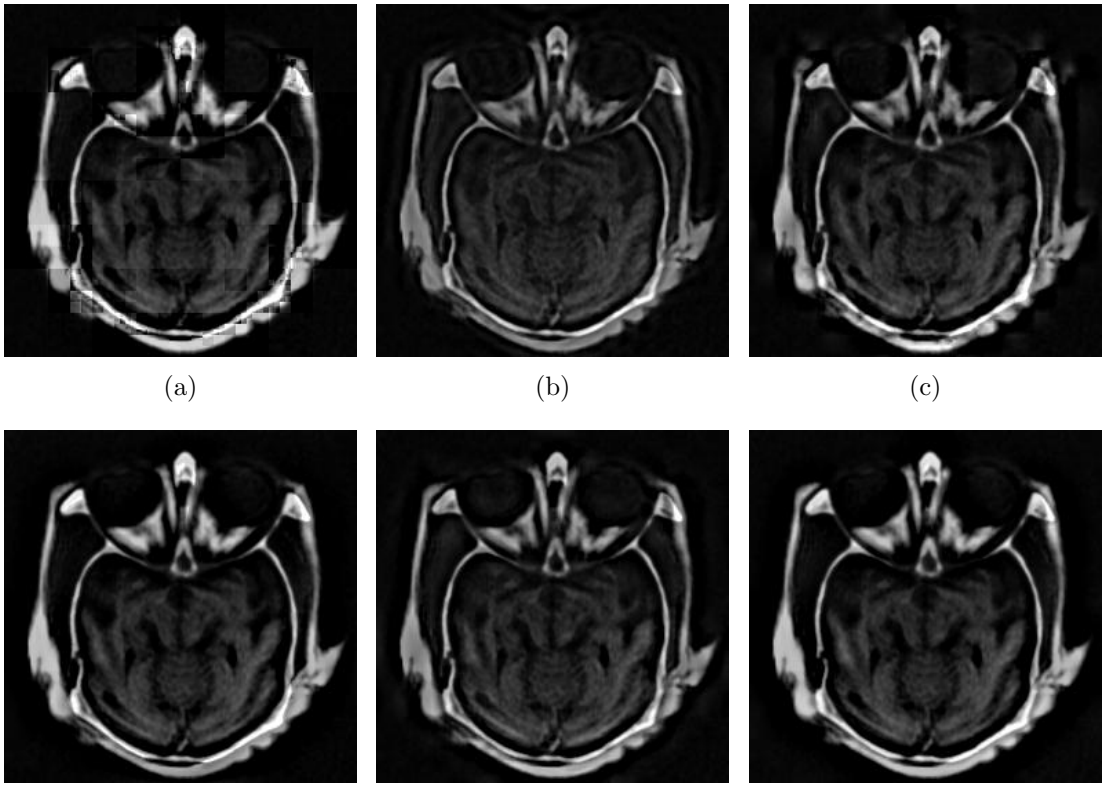
**Figure 3.12:** *Fusion results for the IR-visible image pair of Fig. 3.9(bottom row). (a) DWT fused. (b) CVT fused. (c) ConT fused. (d) UWT fused. (e) DTCWT fused. (f) NSCT fused.*

In order to verify the perceptual accuracy of our findings, Fig. 3.12 shows the obtained fusion results for the IR-visible image pair depicted at the bottom of Fig. 3.9. At first glance no major differences between the depicted images can be found. However, when carefully comparing the source image pair of Fig. 3.9 with the displayed results, we see that Figs. 3.12(a) to 3.12(c) introduce a significant number of artifacts which are not present in any of the source images. These reconstruction errors are especially visible in the immediate vicinity of the roof top, illustrated at the bottom half of the fused images. Although some distortions are also visible in Figs. 3.12(d) to 3.12(f), their appearance is less noticeable, thus indicating the compliance of the calculated fusion scores with subjective perception.

When examining the results listed in Table 3.9 for the medical image fusion scenario, it can again be observed that shift-invariant transforms such as the UWT, DTCWT and the NSCT significantly outperform the DWT, CVT and ConT. How-

| Transform | Filter Bank(s) | | Levels/Directions | $Q_{AB/F}$ | MI | $Q_P$ |
|---|---|---|---|---|---|---|
| DWT | *db1* | | 5 | 0.6625 | **0.3461** | 0.5912 |
| CVT | | | $\{1, 8, 16, 16, 32\}$ | 0.6257 | 0.2251 | 0.5808 |
| ConT | *5/3* | *9/7* | $\{2, 4, 4, 8\}$ | 0.6196 | 0.2500 | 0.5832 |
| UWT | *db1* | | 5 | **0.7302** | 0.2829 | **0.6559** |
| DTCWT | *5/3* | *qshift10-6* | 4 | 0.6611 | 0.2619 | 0.6082 |
| NSCT | *maxflat2* | *dmaxflat7* | $\{4, 8, 8, 16, 16\}$ | 0.7136 | 0.2862 | 0.6479 |

**Table 3.9:** *Global comparison of the best results for the medical image fusion scenario.*



(a) (b) (c)

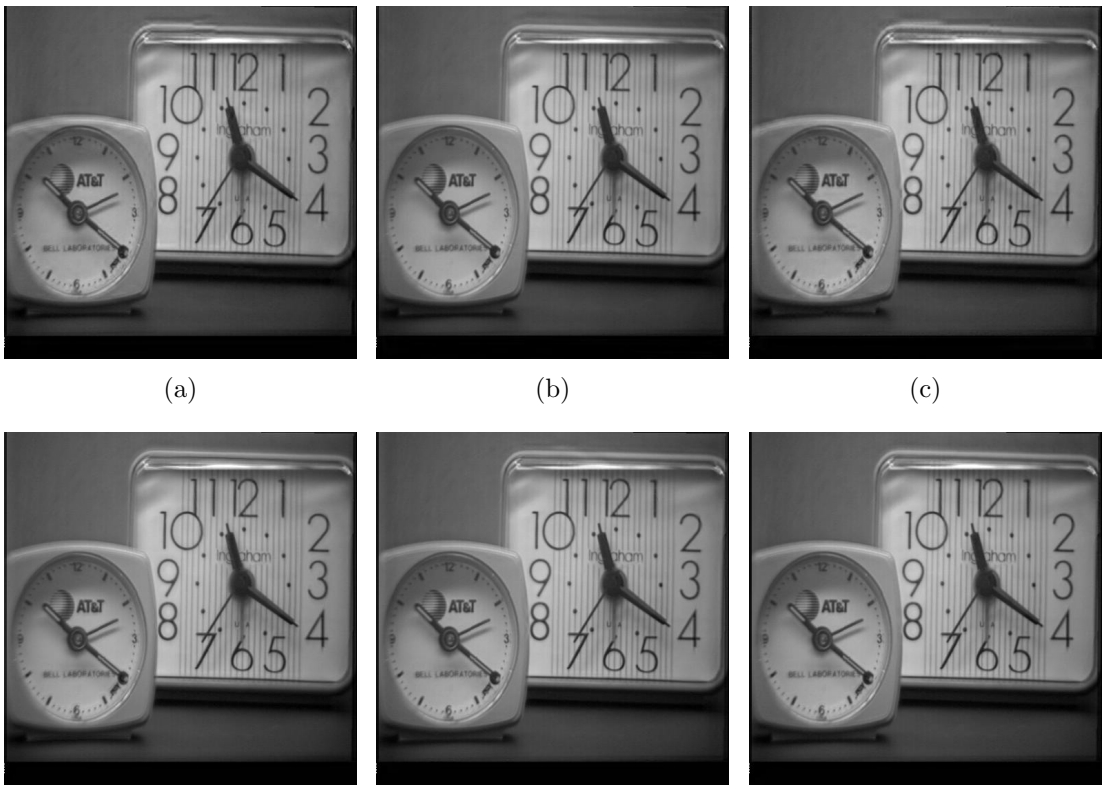
(d) (e) (f)

**Figure 3.13:** *Fusion results for the medical image pair of Fig. 3.10. (a) DWT fused. (b) CVT fused. (c) ConT fused. (d) UWT fused. (e) DTCWT fused. (f) NSCT fused.*

ever, in contrast to the IR-visible fusion scenario, the best performance is achieved for the UWT in combination with the 2-tap Haar filter bank. This is a particular interesting result since it suggests that for some cases, the overall support size of the deployed filter bank appears to have a stronger influence on the overall fusion performance than the number of directional decompositions. We deal with the implications of this assertion thoroughly in the next chapter of this work. In order to perceptually confirm the obtained objective fusion scores, Fig. 3.13 shows the fusion results for the medical source image pair depicted in Fig. 3.10. Indeed, it can be noticed that in the UWT-based fusion case the main features of the source image pair appear to be slightly more accentuated than in all remaining cases, suggesting

| Transform | Filter Bank(s) | | Levels/Directions | $Q_{AB/F}$ | MI | $Q_P$ |
|---|---|---|---|---|---|---|
| DWT | *sym8* | | 5 | 0.6368 | 0.4542 | 0.8663 |
| CVT | | | $\{1, 8, 16, 16, 32, 32\}$ | 0.6649 | 0.4785 | 0.8831 |
| ConT | *5/3* | *9/7* | $\{1, 2, 2, 4\}$ | 0.6350 | 0.4467 | 0.8697 |
| UWT | *db1* | | 4 | **0.6786** | 0.4813 | 0.8804 |
| DTCWT | *5/3* | *qshift10-6* | 5 | 0.6718 | **0.4867** | **0.8865** |
| NSCT | *9/7* | *vk* | $\{1, 2, 2, 4, 4\}$ | 0.6775 | 0.4787 | 0.8838 |

**Table 3.10:** *Global comparison of the best results for the multifocus image fusion scenario.*



(a) (b) (c)

(d) (e) (f)

**Figure 3.14:** *Fusion results for the multifocus image pair of Fig. 3.11. (a) DWT fused. (b) CVT fused. (c) ConT fused. (d) UWT fused. (e) DTCWT fused. (f) NSCT fused.*

the perceptual effectiveness of the employed fusion metrics. Another interesting effect can be observed in Fig. 3.13(a). Here, the number of decomposition levels was chosen too high, resulting in the introduction of blocking artifacts in the final fused image. Apart from the decomposition depth, this effect can also be related to the use of the Haar filter bank which is "notorious" in causing this kind of reconstruction error. Please note that in this case the obtained fusion scores do not reflect the subjective quality of the image. This applies in particular to the MI fusion metric, which seems to confuse the introduced artifacts with important visual information.

The best results for the multifocus fusion example of Fig. 3.11 are listed in Table 3.10. In this scenario the best fusion scores are achieved for the DTCWT,

followed by the UWT and the NSCT. As for the shift-variant transforms, the best fusion performance is achieved using the CVT for all three fusion metrics. Fig. 3.14 shows the corresponding fusion results for all analyzed transforms. Again, at first sight not much difference between the obtained results can be noticed. However, by closer inspection of the fusion results depicted in the top row of Fig. 3.14, it can be observed that some distortions were introduced around the top boarder of the right clock. These artifacts are especially perceivable in Figs. 3.14(a) and 3.14(c) corresponding to the results obtained for the DWT and the ConT. Indeed, by looking at Table 3.10 we note that these two transforms received the worst fusion scores for all three objective metrics.

## 3.4 Conclusions

In this chapter we compared the image fusion performance of six multiscale transforms for two IR-visible, and one medical and multifocus image pair, respectively, using different filter bank settings and decomposition depths. At the time of preparation of this work, the analyzed transforms represented the state-of-the-art in image fusion applications. They mainly differ in their underlying sampling scheme (decimated vs. undecimated) as well as in the offered number of directional decompositions. In all of our simulations the decomposed detail images were fused using the generic "choose max" fusion rule whereas the composite approximation image was computed by simple averaging. The obtained results have been analyzed in terms of the three objective metrics $Q_{AB/F}$, MI and $Q_P$. Additionally, in an attempt to show the perceptual effectiveness of the deployed fusion metrics, the best results for each transform were subject to an informal visual inspection.

The overall comparison performed in this chapter indicated that the best results, regardless of the underlying fusion scenario, can be obtained using redundant, shift-invariant transforms such as the Undecimated Wavelet Transform (UWT), the Nonsubsampled Contourlet Transform (NSCT) and the Dual-Tree Complex Wavelet Transform (DTCWT). We believe that the main advantage of these transforms roots in the offered redundancy, resulting in a higher robustness to rapid changes in coefficient values. This was confirmed by the corresponding fused images which, compared to the set obtained for shift-variant transforms, exhibit a smaller number of decomposition errors and appear more 'natural' to the human eye. This was also reflected in the fusion scores which constantly rate the UWT, DTCWT and the NSCT higher than the Discrete Wavelet Transform (DWT), Curvelet Transform (CVT) and Contourlet Transform (ConT).

Among the shift-invariant transforms, it appears to be difficult to explicitly recommend a single one of them. In fact, whereas in the IR-visible fusion scenario the

NSCT performed best, the UWT and the DTCWT produced the best fusion scores in the medical and multifocus fusion scenarios, respectively. Moreover, a ranking based on the subjective assessment of the obtained results proved to be cumbersome due to the vast similarities between the fused images. As for the overall number of decomposition levels, four to five levels yielded the best results.

Finally, we would like to point out that the results obtained in this chapter are derived from only four different source image pairs, and thus cannot be considered generally valid. However, the main insights gathered in this chapter, namely, the superiority of redundant, shift-invariant transforms as well as the general tendency of multiscale fusion schemes towards filter banks with smaller support sizes, are indeed valuable. In fact, they can be considered as the foundation for the development of a novel image fusion framework, which is the topic of the next chapter.

# Chapter 4

# Multiscale image fusion using the Undecimated Wavelet Transform with spectral factorization and non-orthogonal filter banks

Multiscale transforms are among the most popular techniques in the field of pixel-level image fusion. However, the fusion performance of these methods often deteriorates for images derived from different sensor modalities. In this chapter we demonstrate that for such images, results can be improved using a novel fusion scheme based on the Undecimated Wavelet Transform (UWT) which splits the image decomposition process into two successive filtering operations using spectral factorization of the analysis filters. The actual fusion takes place after convolution with the first filter pair. Its significantly smaller support size leads to the minimization of the unwanted spreading of coefficient values around overlapping image singularities. This usually complicates the feature selection process and may lead to the introduction of reconstruction errors in the fused image. Moreover, we show that the nonsubsampled nature of the UWT allows for the design of non-orthogonal filter banks which are more robust to artifacts introduced during fusion, additionally improving the obtained results. The combination of these approaches leads to a fusion framework which provides clear advantages over traditional multiscale fusion approaches, independent of the underlying fusion rule, and reduces unwanted side effects such as ringing artifacts in the fused image.

The remainder of this chapter is organized as follows. The next section gives an overview on the problem at hand and outlines how spectral factorization can be used to alleviate it. In Section 4.2 the proposed image fusion framework is introduced in detail, whilst Section 4.3 elaborates on the design of non-orthogonal filter banks.

The set of fusion rules used to assess the proposed fusion framework is presented in Section 4.4 of this chapter. Finally, the obtained fusion results are analyzed and compared with other state-of-the-art fusion frameworks in Section 4.5, before we state our main conclusions in Section 4.6.

## 4.1 Motivation

In multiscale pixel-level image fusion, a transform coefficient of an image is associated with a feature if its value is influenced by a feature's pixel. In order to simplify the discussion, we will refer to a given decomposition level $j$, orientation band $p$ and position $m, n$ of a coefficient as its *localization*. A given feature from one of the source images is only conserved correctly in the fused image if all associated coefficients are employed to generate the fused multiscale representation. However, in many situations this is not practical since, given a localization $\mathbf{l}$, the coefficient $y_A(\mathbf{l})$ from image $I_A$ may be associated to a feature $f_A$ and the coefficient $y_B(\mathbf{l})$ from image $I_B$ may be associated to a feature $f_B$. In this case, choosing one coefficient instead of the other may result in the loss of an important salient feature from one of the source images. For example, in the case of a camouflaged person hiding behind a bush the person may appear only in the infrared (IR) image and the bush only in the visible image. If the bush has high textural content, this may result in large coefficient values at coincident localizations in both decompositions of an IR-visible image pair. However, in order to conserve as much as possible of the information from the scene, most coefficients belonging to the person (IR image) and the bush (visible image) would have to be transferred to the fused decomposition. If there are many such coefficients at coincident localizations, a fusion rule that chooses just one of the coefficients for each localization may introduce discontinuities in the fused subband signals. These may lead to reconstruction errors such as ringing artifacts or substantial loss of information in the final fused image.

It is important to note that the above mentioned problem is aggravated with the increase of the support of the filters used during the decomposition process. This results in an undesirable spreading of coefficient values over the neighborhood of salient features, introducing additional areas that exhibit coefficients in the source images with coincident localizations. In a previous work, Petrović and Xydeas dealt with this problem by employing image gradients [41]. In this work we propose a novel UWT-based pixel-level image fusion approach, which attempts to circumvent the coefficient spreading problem by splitting the image decomposition procedure into two successive filter operations using spectral factorization of the analysis filters. A schematic flow-chart of the suggested image fusion framework is given in Fig. 4.1. The co-registered source images are first transformed to the UWT domain by using

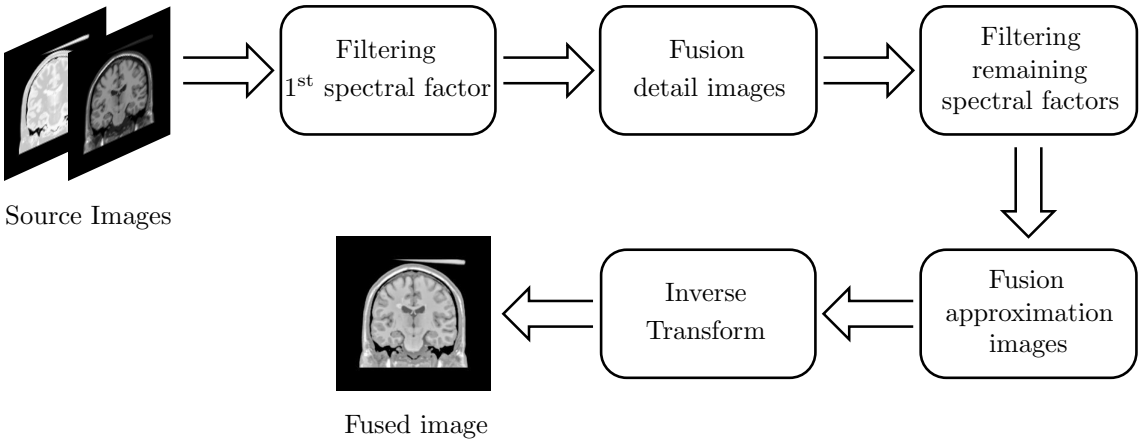



**Figure 4.1:** *Schematic diagram of the proposed UWT-based fusion framework with spectral factorization.*

a very short filter pair, derived from the first spectral factor of the overall analysis filter bank. After the fusion of the high-pass coefficients, the second filter pair, consisting of all remaining spectral factors, is applied to the approximation and fused, detail images. This yields the first decomposition level of the proposed fusion approach. Next, the process is recursively applied to the approximation images until the desired decomposition depth is reached. After merging the approximation images at the coarsest scale the inverse transform is applied to the composite UWT representation, resulting in the final fused image.

Notice that this methodology is in contrast to conventional multiscale image fusion approaches, where the detail image fusion is not performed until the input image signals are fully decomposed using an analysis filter bank without spectral factorization. In addition, the implemented filter banks were especially designed for the use with the UWT and exhibit useful properties such as being robust to the ringing artifact problem. In the course of this chapter, we show that our framework significantly improves fusion results for a large group of input images.

In fact, one may also use spatial domain techniques as briefly introduced in Section 1.3.2 to avoid the problems associated with coincident coefficient localizations. However, these approaches face other difficulties which may limit their practicability in certain situations. Generally speaking, the biggest challenge for spatial domain fusion techniques is the question on how to measure saliency within an image - a problem which can be solved more easily using transform-based approaches, due to the frequency and/or orientation selectivity provided by them.

Note that in order to simplify the discussion, we assume, without loss of generality, that the fused image is generated from two source images $I_A$ and $I_B$ which are assumed to be registered, as discussed in Chapter 1. Consequently, no image registration technique is applied prior to the fusion process. Moreover, in the remainder of this chapter we resort to the same notation as given in Section 2.1.

## 4.2 The UWT-based fusion scheme with Spectral Factorization

Plenty of transforms are at our disposal to perform image fusion tasks, among them the Discrete Wavelet Transform (DWT), the Curvelet Transform (CVT) and the Contourlet Transform (ConT), as well as the UWT, the Dual-Tree Complex Wavelet Transform (DTCWT) and the Nonsubsampled Contourlet Transform (NSCT). A first classification can be made based on the underlying redundancy and shift-variance of these transforms. Whereas the highly redundant UWT, DTCWT and NSCT are invariant to shifts occurring in the input images, the DWT, CVT and CT represent shift-variant transforms with no or limited redundancy. As shown in the previous section, redundancy and shift-invariance are desirable properties in image fusion since they allow for a higher robustness to rapid changes in coefficient values, thus reducing the amount of reconstruction errors in the fused image. This was also acknowledged in various studies such as [2], [5] and [23], among others. Motivated by these observations, we opt to discard the DWT, CVT and CT and focus solely on redundant transforms in our ongoing discussion.

Another crucial point in multiscale pixel-level image fusion frameworks is the choice of an appropriate filter bank. Most research work do not focus on this issue but simply state that filters with small support produce better results. Fig. 4.2 attempts to illustrate the impact of the length of the chosen filter bank on the fusion performance. In this example the high-pass portions of two 1-D step functions are fused using one stage of the 2-tap Haar and 6-tap Daubechies 'db3' filters, respectively. The applied fusion rule is a very simple "choose max" fusion rule. The high-pass subbands, obtained by applying the Haar filter, can be seen in Figs. 4.2(b) and (e), whereas the result using the 6-tap 'db3' filter is illustrated in Figs. 4.2(c) and (f). It can be observed that the 'db3' filter needs five coefficients to represent the step change. Thus, although most energy is concentrated in the central coefficient, the remaining four coefficients correspond to regions where no change in the signal value occurred. When attempting to fuse the two 'db3' filtered high-pass subbands we are confronted with a problem, namely, to combine the two signals without losing information. This can be observed in Fig. 4.2(h), where not all non-zero coefficients from Figs. 4.2(c) and (f) could be incorporated. On the other hand, the Haar filtered signal contains only one non-zero coefficient corresponding exactly to the position of the signal transition. Thus, as illustrated in Fig. 4.2(g), both non-zero coefficients are transferred to the fused image without any loss of information. Therefore, it can be concluded that filters with large support size may result in an undesirable spreading of coefficient values which, in case of salient features located very close to each other in both input images, may lead to coefficients with coincident localizations

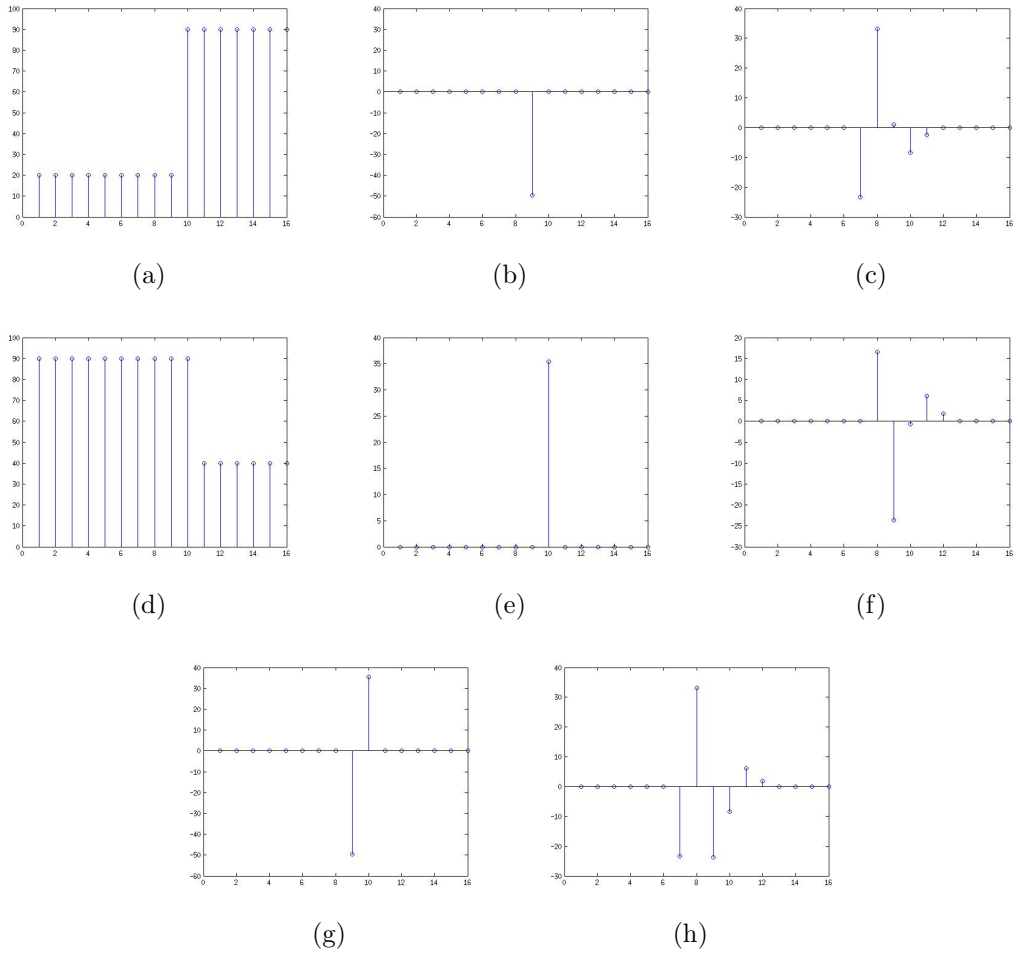


**Figure 4.2:** *Coefficient spreading effect. (a) and (d) Input signals. (b) and (e) Haar filtered input signals. (c) and (f) 'db3' filtered input signals. (g) Fusion of the Haar filtered signals. (h) Fusion of the 'db3' filtered signals.*

in the transform domain. Since it is difficult to resolve such overlaps, distortions may be introduced during the fusion process, such as ringing artifacts or even loss of information.

Although the situation depicted in Fig. 4.2 may seem at first somewhat artificial, we will see in the remainder of this chapter that multisensor images and, among them, especially medical image pairs often exhibit similar properties. Hence, for these images the fusion performance considerably degrades with an increase of the filter size. We can therefore reduce the problem of choosing a proper redundant multiscale transform to its ability to incorporate a filter bank with a sufficiently small support size, thus minimizing the coefficient spreading problem. From this point of view, the UWT appears to be an attractive choice, since due to the standard tensor product construction in 2-D, it offers directionality without increasing the overall length of the implemented filter bank - a property not shared by the NSCT and DTCWT. As for the NSCT, the increased filter lengths are mainly due to the

iterated nature of the nonsubsampled directional filter bank involved. In particular, for every increase in number of directions by a power of two, another filter bank level needs to be added (see Sections 3.1.3 and 3.1.6 for a more thorough discussion on the construction of directional filter banks). Thus, the combined support of the filters within one particular filter bank branch is equivalent to the convolution of all individual filters within the respective branch. In the case of the DTCWT, as reported in [96], the increased filter length is due to the half-sample delay condition of eq. (3.36) imposed on the filter banks involved, which results in longer filters than in the real wavelet transform case. From the above, we can conclude that, even though the NSCT and the DTCWT possess useful properties, such as their ability to incorporate a higher number of orientations, they are, in general, less suited to implement filter banks with a small support size.

Following the remarks stated so far, we are tempted to arrive at the conclusion that the best fusion results for source images derived from different sensor modalities are obtained by simply applying the UWT in combination with the very short 2-tap Haar filter bank. Indeed, surprisingly good results are achieved using this simple fusion strategy for IR-visible and medical image fusion, as demonstrated in Chapter 3. However, the Haar filter bank presents some well-known deficiencies, like the introduction of blocking artifacts when reconstructing an image after manipulation of its wavelet coefficients, which might deteriorate the fusion performance in certain situations. This is mainly due to the lack of regularity exhibited by the Haar wavelet [94]. Roughly speaking, the regularity of a wavelet or scaling function ($\psi(t)$ and $\phi(t)$, respectively) relates to the number of continuous derivatives that a wavelet has. In case of the Haar wavelet, the low-pass analysis filter, $H(z)$, has only one zero at $z = -1$, leading to the well-known, non-smooth Haar scaling function. In order to construct smoother scaling functions, more zeros have to be introduced at $z = -1$, inevitably leading to filters with longer support [81]. Actually, for the orthogonal case the Daubechies wavelets discussed so far are optimal in this sense since they have a minimum support size for a given regularity. Fig. 4.3 shows the scaling and wavelet functions of the Haar and the 8-tap 'db4' wavelet, with four zeros at $z = -1$. As expected, the 'db4' wavelet depicts smooth scaling and wavelet functions. Please note that, in the case of (bi)orthogonal wavelets, the regularity of the scaling function is similar to the regularity of the wavelet function. Both are closely related to their number of vanishing moments. Apart from its impact on the smoothness of the corresponding wavelet functions, the regularity is also a measure of the flatness of the scaling and wavelet function around $\omega = 0$ and $\omega = \pi$ in the frequency domain, respectively. This effect is illustrated in Fig. 4.4, where the frequency responses of the Haar, 'db4' and 'db10' wavelets are shown.

Based on these observations we arrive at the following question: How can we

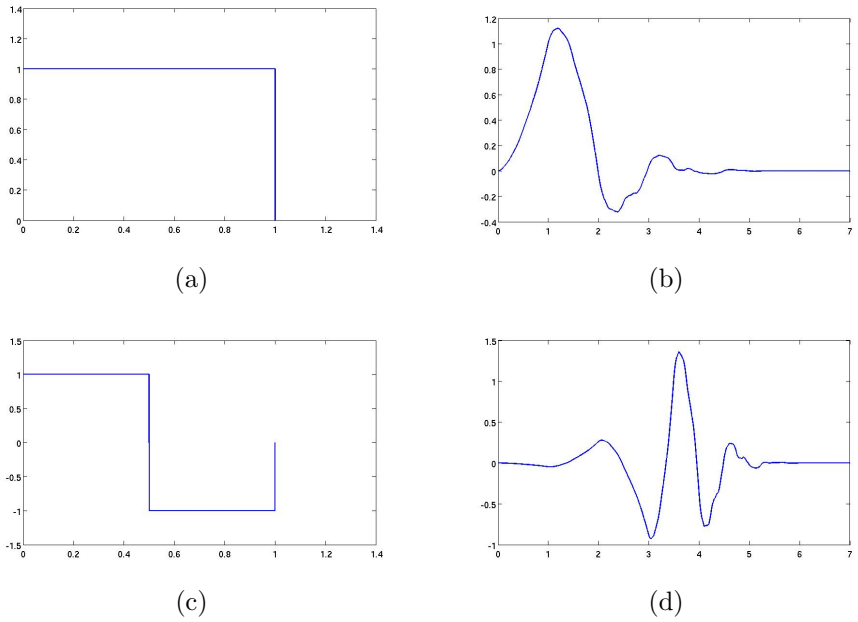


**Figure 4.3:** *(a) Haar scaling function. (b) 'db4' scaling function. (c) Haar wavelet function. (d) 'db4' wavelet function.*

combine the advantages of filters with small support size with the ones of filter banks exhibiting a high degree of regularity in the context of image fusion? In conventional multiscale fusion approaches this dilemma usually results in a trade-off between short-length filters and filters with higher regularity and better behavior in the frequency domain, usually with a small bias towards filter banks with short support sizes. In this paper we propose a novel UWT-based fusion approach that splits the filtering process into two successive filtering operations and performs the actual fusion after convolving the input signal with the first filter pair, exhibiting a significantly smaller support size than the original filter. The proposed method is based on the fact that the low-pass analysis filter $H(z)$ and the corresponding high-pass analysis filter $G(z)$ can always be expressed in the form

$$\begin{aligned} H(z) &= (1+z^{-1})P(z) \\ G(z) &= (1-z^{-1})Q(z) \end{aligned} \tag{4.1}$$

by spectral factorization in the $z$-transform domain. This can be inferred from the regularity and the admissibility condition which state that the filters $H(z)$ and $G(z)$ within an undecimated, perfect-reconstruction filter bank have to have at least one zero at $z = -1$ and $z = 1$, respectively. The interested reader will find more information on this topic in e.g. [81], [95] and [112].

In our framework the input images are first decomposed by applying a Haar filter pair, represented by the first spectral factors $(1 + z^{-1})$ and $(1 - z^{-1})$, respectively. The resulting horizontal, vertical and diagonal detail images can afterwards be fused

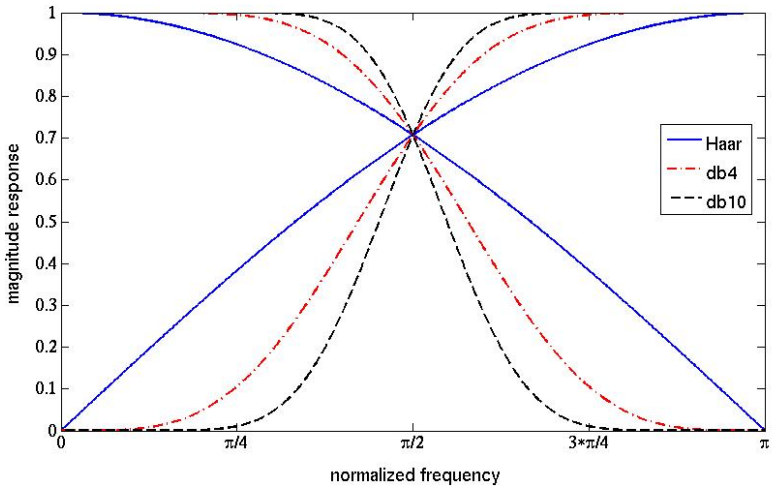



**Figure 4.4:** *Frequency response of the Haar, 'db4' and 'db10' scaling and wavelet functions.*

according to an arbitrary fusion rule. Next, the filter pair represented by the second spectral factor ($P(z)$ and $Q(z)$ in eq. (4.1)), is applied to the approximation and fused detail images, yielding the first decomposition level of the proposed fusion scheme. For each subsequent level, the analysis filters are upsampled according to the "à trous" algorithm, leading to the following, generalized analysis filter bank

$$\begin{aligned} H(z^{2^{j-1}}) &= (1+z^{-2^{j-1}})P(z^{2^{j-1}}) \\ G(z^{2^{j-1}}) &= (1-z^{-2^{j-1}})Q(z^{2^{j-1}}) \end{aligned} \tag{4.2}$$

and the aforementioned procedure is recursively applied to the approximation images, until the desired number of decomposition levels is reached. After merging the low-pass approximation images, the final fused image is obtained by applying the inverse transform, using the corresponding synthesis filter bank without spectral factorization.

The implementation of the proposed algorithm for two 1-D signals $x_A$ and $x_B$ and two decomposition levels is depicted in Fig. 4.5, where **F** symbolizes the fusion of the high-pass coefficients. It is important to stress that spectral factorization is not applied to the low-pass filter $H(z)$ since it is assumed that all salient features of the input signals are embodied in the high-frequency coefficients. Although this assumption remains also true for images, when using separable filters the horizontal and vertical detail bands are obtained by applying both low-pass and high-pass filters to the columns and rows of the input images. Thus, it is necessary to apply spectral factorization also to the low-pass channel. Only in case of the low-low channel (successive application of $H(z)$ to the columns and rows of the input images) spectral factorization is not employed. The implementation of the first stage of our image fusion framework is depicted in Fig. 4.6.

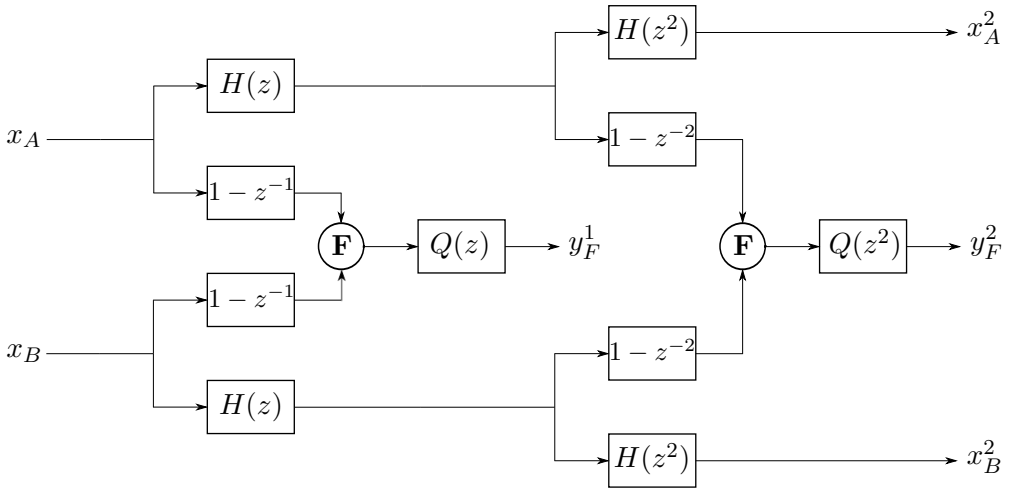



**Figure 4.5:** *Implementation of the UWT-based fusion scheme with spectral factorization for two decomposition levels in 1-D.*

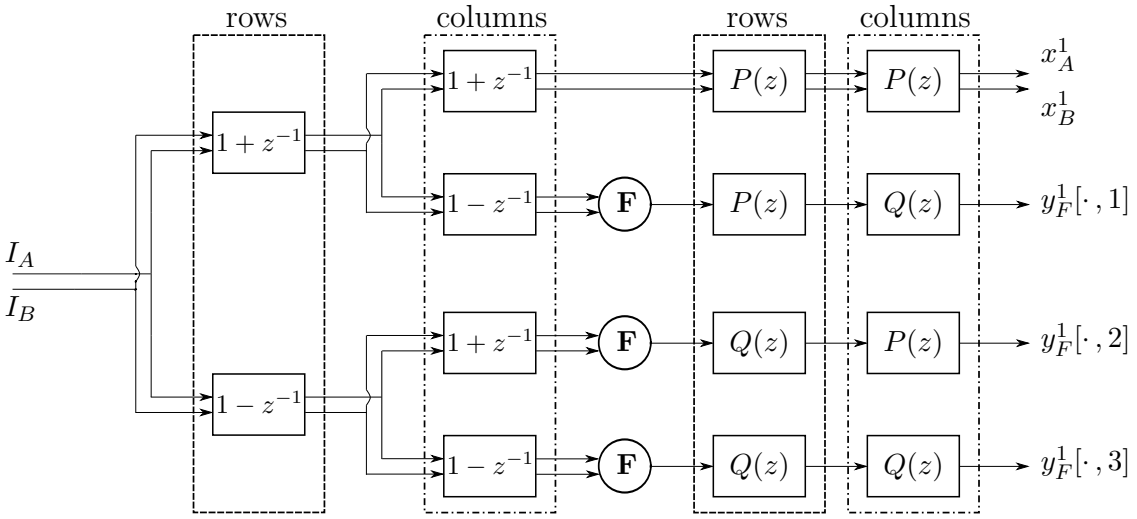



**Figure 4.6:** *Implementation of the* $1^{st}$ *stage of the UWT-based fusion scheme with spectral factorization.*

It is worth mentioning that the upsampling strategy given in eq. (4.2) is not the only possible choice. Especially, if we note that $(1 - z^{-2^{j-1}})$ can be further decomposed as

$$(1 - z^{-2^{j-1}}) = (1 - z^{-1}) \cdot (1 + z^{-1}) \cdot (1 + z^{-2}) \cdot \ \ldots \ \cdot (1 + z^{-2^{j-2}}) \tag{4.3}$$

the upsampled high-pass filter $G(z^{2^{j-1}})$ can be factorized as

$$G(z^{2^{j-1}}) = (1 - z^{-1})R_j(z)Q(z^{2^{j-1}}) = (1 - z^{-1})Q_j(z), \tag{4.4}$$

where $R_j(z)$ consists of all but the first factor of eq. (4.3). Hence, for all scales the detail bands can be fused after applying the same, non-upsampled Haar filter from the first stage. However, experiments showed that, in general, the results change

only marginally, compared to the original filter setup. Thus, we only work with the upsampling strategy given by eq. (4.2).

The novelty of the proposed fusion framework lies in its ability to combine the properties of filters with short support size with filters with large support size and therefore higher regularity. In more detail, due to the compact support of the used $(1 \pm z^{-2^{j-1}})$ factors the undesirable spreading of coefficient values in the neighborhood of salient features during the convolution process is largely reduced. This allows for a more reliable feature selection and reduces both the introduction of distortions and the loss of contrast information during the fusion process, conditions commonly observed in traditional multiscale fusion frameworks. The subsequent filtering with the second spectral factor accounts for the freedom of implementing an arbitrary filter bank (satisfying the perfect reconstruction condition), hence combining the advantages of a very short filter with the benefits of filters with higher orders. In other words, we avoid the introduction of blocking artifacts during reconstruction, as well as the coefficient spreading problem. The proposed fusion framework differs from conventional multiscale fusion methods, where the actual fusion is only applied after fully decomposing the input images using an analysis filter pair without spectral factorization. Furthermore, note that the spectral factorization scheme, as presented in this subsection, cannot be straightforwardly adapted to the NSCT and the DTCWT. This is mainly due to the filter design restrictions imposed by these transforms, preventing the meaningful application of such a factorization scheme. As we are going to show in Section 4.5, the presented approach is particularly well suited for the fusion of IR-visible and medical images, which tend to exhibit a high degree of information at coincident localizations. For these image groups the presented framework outperforms traditional fusion frameworks based on the DTCWT and NSCT.

In the next section a new class of filters, which has not been used in the context of image fusion previously, is introduced. In more detail, we place our emphasis on non-orthogonal filter banks which do not satisfy the anti-aliasing condition of the DWT and can therefore only be used in the nonsubsampled case. We will see that this lack of orthogonality allows for the implementation of filter banks with useful properties such as being more robust to ringing artifacts.

## 4.3 Filter bank design

Due to the nonsubsampled nature of the UWT, many ways exist to construct the fused image from its wavelet coefficients. For a given analysis filter bank $(h, g)$, any synthesis filter bank $(\tilde{h}, \tilde{g})$ satisfying the perfect reconstruction condition of eq. (3.8) can be used for reconstruction. This is considerably simpler and offers

more design freedom than in the decimated case, where the anti-aliasing condition of eq. (3.9) has to be obeyed as well, imposing considerable constraints on the filter bank design. As a consequence, filter banks can be used such that $(\tilde{h}, \tilde{g})$ are positive, making the reconstruction more robust to ringing artifacts. In the remainder of this section these filters, which are later used in our experiments, are explained in more detail. A more thorough discussion on filter bank design for undecimated wavelet decompositions can be found in [94] and [112]. Again, we would like to point out that none of these filters obey the anti-aliasing condition and can therefore only be used in the undecimated case.

We start our discussion with the Isotropic Undecimated Wavelet Transform of Section 3.1.4, which is frequently used in multispectral image fusion. In this approach, only one detail image for each scale is obtained and not three as in the general case. It is implemented using the non-orthogonal, 1-D filter bank

$$\begin{array}{lllll} h[n] & & & = & \dfrac{[1,4,6,4,1]}{16} \\ g[n] & = & \delta[n] - h[n] & = & \dfrac{[-1,-4,10,-4,-1]}{16}\ , \\ \tilde{h}[n] & = & \tilde{g}[n] & = & [0,0,1,0,0] \end{array} \tag{4.5}$$

where $h$ is derived from the $B_3$-spline function. Please note that, by choosing $g[n] = \delta[n] - h[n]$ any low-pass filter $h$, having at least one zero at $z = -1$, can be used. However, $B$-spline functions have some remarkable properties which make them very good choices for wavelet analysis. For example, if we recall that $H(z)$ can be factorized as $H(z) = (1 + z^{-1})^L \cdot P(z)$, it can be shown that the $B$-spline function of degree $n = L - 1$ is the shortest and most regular scaling function of order $L$, with $P(z) = 1$ [113].

The standard three-directional UWT can be obtained by expanding the filter bank to 2-D as described in eqs. (3.10) to (3.13), leading to the following representation of the original image

$$I[n,m] = x^J[n,m] + \sum_{j=1}^{J} \sum_{d=1}^{3} y^j[n,m,d], \tag{4.6}$$

which is conceptually very close to the reconstruction given in eq. (3.30).

This approach has some interesting characteristics. For example, due to the lack of convolutions during reconstruction, no additional distortions are introduced when constructing the fused image. Furthermore, since the fused image is obtained by a simple co-addition of all detail images and the approximation image, a very fast reconstruction is possible. On the other hand, distortions introduced during the fusion process remain unfiltered in the reconstructed image.

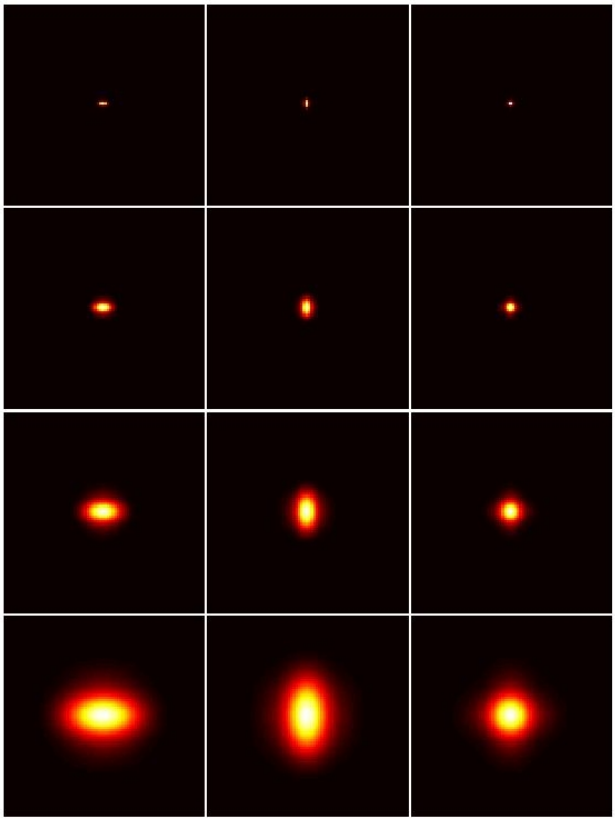


**Figure 4.7:** *Backprojection of a single wavelet coefficient at different scales and directions for the filter bank given in eq. (4.7). From left to right, the coefficient belongs to the horizontal, vertical and diagonal bands. From top to bottom, the scale increases from one to four. Each scale and direction has been normalized such that it occupies the full dynamic range.*

Alternatively, if we choose $h$ and $g$ as before but define the synthesis low-pass filter $\tilde{h}$ as $h$, we obtain a filter $\tilde{g}$ given by $\tilde{g} = \delta + h$. This yields filters with the following coefficients

$$\begin{aligned} h[n] &= \tilde{h}[n] &&= \frac{[1,4,6,4,1]}{16} \\ g[n] &= \delta[n] - h[n] &&= \frac{[-1,-4,10,-4,-1]}{16} \\ \tilde{g}[n] &= \delta[n] + h[n] &&= \frac{[1,4,22,4,1]}{16} \end{aligned} \quad . \tag{4.7}$$

In this scenario $\tilde{g}$ consists entirely of positive coefficients, being thus no longer related to a wavelet function. On the other hand, such a lack of oscillations provides a reconstruction less vulnerable to ringing artifacts. Additionally, distortions introduced during the fusion stage are not transferred unprocessed to the reconstructed image as in the standard case where only summations are involved during reconstruction. Fig. 4.7 shows the backprojection of a wavelet coefficient at different scales and directions for the filter bank given in eq. (4.7). Note that each scale and

direction has been normalized such that the full dynamic range is occupied. It can be observed that all images solely exhibit positive values.

A slight variation of the previous example is obtained by defining $g = \delta - h * h$, resulting in the filter bank

$$\begin{array}{rclcl} h[n] & = & \tilde{h}[n] & = & \dfrac{[1,2,1]}{4} \\ g[n] & = & \delta[n] - h[n] * h[n] & = & \dfrac{[-1,-4,10,-4,-1]}{16} \\ \tilde{g}[n] & = & \delta[n] & = & [0,0,1,0,0] \end{array}, \tag{4.8}$$

where $h$ is derived from the $B_1$-spline function. Please note that for the same choice of $h$ the last two approaches are conceptually similar, since $(\delta[n] - h[n]) * (\delta[n] + h[n]) = \delta[n] * \delta[n] + h[n] - h[n] - h[n] * h[n] = \delta[n] - h[n] * h[n]$. Thus, the analysis high-pass filter of eq. (4.8) can be attained by a convolution of the analysis and synthesis high-pass filters of eq. (4.7).

Finally, we would like to point out that plenty of other alternatives exist. For example the filter bank

$$\begin{array}{rclcrcl} h[n] & = & \dfrac{[1,1]}{2} & & g[n] & = & \dfrac{[-1,2,-1]}{4} \\ \tilde{h}[n] & = & \dfrac{[1,3,3,1]}{8} & & \tilde{g}[n] & = & \dfrac{[1,6,1]}{4} \end{array}, \tag{4.9}$$

also leads to a solution where both synthesis filters are positive.

We will see that these filters, in combination with spectral factorization, yield superior fusion results compared to traditional techniques.

## 4.4 Fusion rules

As shown in Chapter 2, a wide range of combination schemes can be found in the literature to fuse an arbitrary input image pair. In general, these rules vary greatly in terms of their complexity and effectiveness. The spectral factorization method proposed in this chapter can be employed together with any fusion rule. Therefore, in order to assess the effectiveness of the proposed method, we applied four different fusion rules.

The first investigated combination scheme is the simple 'choose max' (CM) or maximum selection fusion rule given in eq. (3.51), where the coefficient yielding the highest energy is directly transferred to the fused decomposed representation.

However, even though the CM fusion rules have been shown to be effective, they do not take into account that, by construction, each coefficient within a multiscale

decomposition is related to a set of coefficients in other orientation bands and decomposition levels, as schematically demonstrated in Fig. 2.5 for the case of the DWT. Thus, in order to conserve a given feature from one of the source images, all the coefficients corresponding to it have to be transferred to the composite multiscale representation as well. One way to improve the fusion results is therefore the use of intra-scale grouping (see Section 2.3.4) in combination with the CM fusion scheme of eq. (3.51) (CM-IS). By this rule, at each location $\mathbf{n}$, the fused, detail images $y_F^j$ are defined as

$$y_F^j[\mathbf{n},p] = \begin{cases} y_A^j[\mathbf{n},p] & \text{if } \sum_{q=1}^{Q} \left|y_A^j[\mathbf{n},q]\right| > \sum_{q=1}^{Q} \left|y_B^j[\mathbf{n},q]\right| \\ y_B^j[\mathbf{n},p] & \text{otherwise} \end{cases}, \tag{4.10}$$

where the fusion decision at each decomposition level $j$ is taken jointly for all orientation bands $p$.

Since the combination schemes of eqs. (3.51) and (4.10) suffer from a relative low tolerance against noise which may lead to a "salt and pepper" appearance of the selection maps, robustness can be added to the fusion process using an area-based selection criteria [4]. For this purpose we expand the CM-IS combination scheme of eq. (4.10) by defining the following fusion rule (CM-A): Calculate the activity $a_k^j$ of each coefficient as the energy within a $3\times3$ window centered at the current coefficient position as given in eq. (2.3) and select the coefficient which yields the highest activity, again, by considering the intra-scale dependencies between coefficients from different orientation bands

$$y_F^j[\mathbf{n},p] = \begin{cases} y_A^j[\mathbf{n},p] & \text{if } \sum_{q=1}^{Q} a_A^j[\mathbf{n},q] > \sum_{q=1}^{Q} a_B^j[\mathbf{n},q] \\ y_B^j[\mathbf{n},p] & \text{otherwise} \end{cases}. \tag{4.11}$$

The fusion rules discussed so far work well under the assumption that only one of the source images provides the most useful information. However, this is not always valid and a fusion rule which uses a weighted combination of the transform coefficients may give better results. Following this reasoning we implement as the fourth fusion rule, a modified version of the one given by Burt and Kolczynski in [32] (CM-AM). As in the original implementation, we start by calculating the activity $a_k^j$ as given in eq. (2.3) as well as the match measure $m_{AB}^j$ of eq. (2.4), for a window of size $3 \times 3$. The fused coefficients are then obtained by weighted averaging as demonstrated in eq. (2.5). The fusion weights $w_A^j$ and $w_B^j$ are determined by using

a slightly altered version of eq. (2.6) such that

$$w_A^j[\mathbf{n},p] = \begin{cases} 1 & \text{if } m_{AB}^j[\mathbf{n},p] \leq T \text{ and } \sum_{q=1}^{Q} a_A^j[\mathbf{n},q] > \sum_{q=1}^{Q} a_B^j[\mathbf{n},q] \\ 0 & \text{if } m_{AB}^j[\mathbf{n},p] \leq T \text{ and } \sum_{q=1}^{Q} a_A^j[\mathbf{n},q] \leq \sum_{q=1}^{Q} a_B^j[\mathbf{n},q] \\ \frac{1}{2} + \frac{1}{2}\left(\frac{1-m_{AB}^j[\mathbf{n},p]}{1-T}\right) & \text{if } m_{AB}^j[\mathbf{n},p] > T \text{ and } \sum_{q=1}^{Q} a_A^j[\mathbf{n},q] > \sum_{q=1}^{Q} a_B^j[\mathbf{n},q] \\ \frac{1}{2} - \frac{1}{2}\left(\frac{1-m_{AB}^j[\mathbf{n},p]}{1-T}\right) & \text{if } m_{AB}^j[\mathbf{n},p] > T \text{ and } \sum_{q=1}^{Q} a_A^j[\mathbf{n},q] \leq \sum_{q=1}^{Q} a_B^j[\mathbf{n},q] \end{cases} \tag{4.12a}$$

$$w_B^j[\mathbf{n},p] = 1 - w_A^j[\mathbf{n},p] \tag{4.12b}$$

for some threshold $T$, where we ensure that the fusion decision is taken jointly for all directional decompositions.

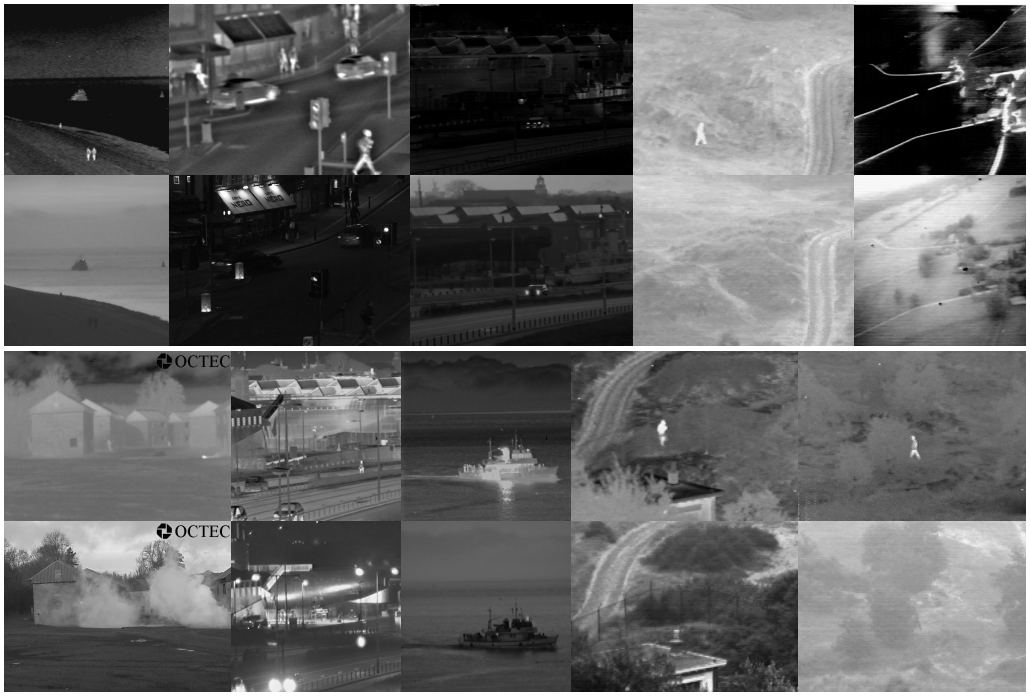
Additionally, since our proposed fusion framework does not suggest any improvements regarding the fusion of the approximation images, for all previously discussed combination schemes the composite approximation coefficients are obtained by averaging as stated in eq. (3.52).

## 4.5 Results

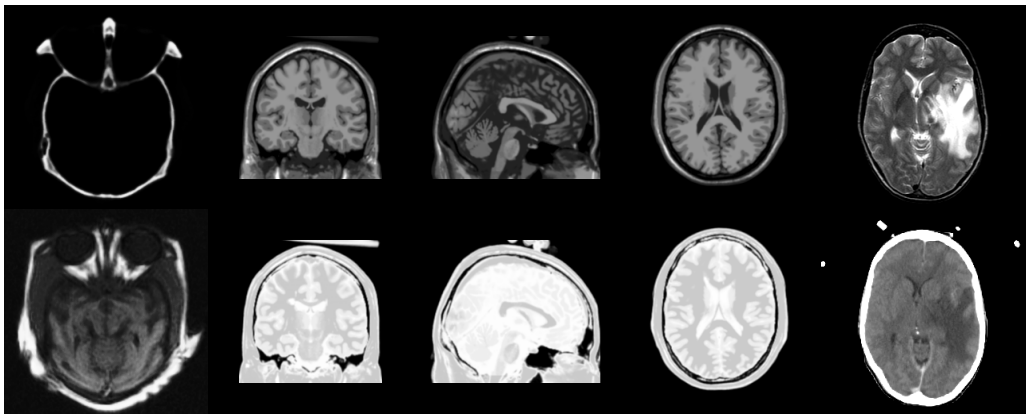
In this section the performance of the proposed fusion framework with spectral factorization is investigated, using three different sets of image pairs. The first set consists solely of IR-visible image pairs, whereas the second and third group comprise medical and multifocus images, respectively. The corresponding thumbnails of all used source images, divided into their corresponding groups, are illustrated in Fig. 4.8.

The performance of the proposed UWT fusion scheme with spectral factorization is compared to the results obtained by applying the NSCT, the DTCWT and the UWT without spectral factorization. As for the NSCT and the DTCWT, we followed the recommendations published in [2] regarding the filter choices and (in case of the NSCT) number of directions. Table 4.1 lists the used settings for the NSCT and DTCWT for each image group. Please note that the chosen filter names correspond to the ones used in Chapter 3.

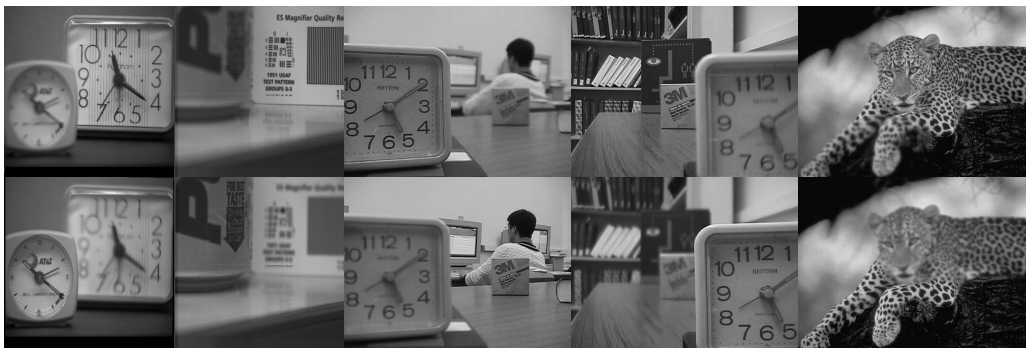
In case of the UWT-based image fusion, we mainly concentrate on the filters from Section 4.3. Hence, in our experiments the non-orthogonal filter banks from eqs. (4.5), (4.7), (4.8) and (4.9) are used. Additionally, we also consider some




(a)



(b)



(c)

**Figure 4.8:** *Thumbnails of all image pairs used for evaluation purposes. (a) IR-visible images (ten pairs). Top row consists of IR images, whereas the bottom row represents the corresponding visible images. (b) Medical images (five pairs). (c) Multifocus images (five pairs).*

| Image Class | Transform | Filters | | Directions |
|---|---|---|---|---|
| Infrared-visible | NSCT | *maxflat3* | *9/7* | $\{4, 8, 8, 16\}$ |
| | DTCWT | *5/3* | *qshift10-6* | |
| Medical | NSCT | *maxflat3* | *vk* | $\{4, 8, 8, 16\}$ |
| | DTCWT | *5/3* | *qshift10-6* | |
| Multifocus | NSCT | *9/7* | *9/7* | $\{4, 8, 8, 16\}$ |
| | DTCWT | *near_sym_a* | *qshift10-6* | |

**Table 4.1:** *Transform settings for the NSCT and DTCWT (according to [2]). The NSCT filter banks to the left (third column) are applied during the nonsubsampled pyramidal decomposition stage whereas the filter banks on the right side (fourth column) are used within the nonsubsampled directional decomposition. The number of directional decompositions, in increasing order from the* $1^{st}$ *to the* $4^{th}$ *stage, is given in the last column. As for the DTCWT, the filter banks to the left are employed in the first decomposition stage whereas the filter banks on the right hand side are applied in all remaining stages.*

biorthogonal filters, which are frequently used in image processing applications such as the CDF 5/3, CDF 9/7 and Rod 6/6 filter bank [114]. In order to avoid referring to filter banks by their respective equation numbers, we associate the following names to them. Henceforth, the filter banks presented in eqs. (4.5), (4.7) and (4.8) are referred to as 'Spline_1', 'Spline_2' and 'Spline_3' filter banks, respectively. The filter bank given in eq. (4.9) will be called 'Haar_1' filter bank, since $h$ is deduced from the Haar low-pass filter. Please note that, in case of the NSCT and DTCWT, different filter banks have been used for each of the three classes of input images, according to Table 4.1. In contrast, for the UWT-based approaches, the same filter banks are used for all three image classes. For all transforms four decomposition levels are chosen. As for the objective evaluation of the achieved results we use the $Q_{AB/F}$, $Q_P$ and MI fusion metrics as described in Section 3.2.

Tables 4.2, 4.3 and 4.4 list the average results as well as the corresponding standard deviations ($\sigma$) for all infrared-visible, medical and multifocus image pairs, respectively, obtained by applying the DTCWT, NSCT and UWT with and without spectral factorization. In all of these simulations the low-pass approximation images are fused using the averaging operation given in eq. (3.52), whereas the fused detail images are obtained by applying the "choose max" (CM) fusion rule of eq. (3.51). It can be noted that the proposed spectral factorization method works well for IR-visible and medical image pairs, but does not yield any improvements for multifocus image pairs. In a nutshell, this is due to the fact that multifocus image pairs only differ in their high frequency content but are identical otherwise. Thus, the source images tend not to contain salient features at coincident localizations. Therefore, a situation as depicted in Fig. 4.2, where the effect of the coefficient spreading problem for two 1-D step functions is shown, is unlikely to occur. Consequently,

| Transform | $Q_{AB/F}$ | | MI | | $Q_P$ | |
|---|---|---|---|---|---|---|
| | Mean | $\sigma$ | Mean | $\sigma$ | Mean | $\sigma$ |
| DTCWT | 0.5664 | 0.0709 | 0.1538 | 0.0510 | 0.7707 | 0.0465 |
| NSCT | **0.5786** | **0.0737** | **0.1563** | **0.0525** | **0.7719** | **0.0489** |

(a)

| Filter Bank | $Q_{AB/F}$ | | MI | | $Q_P$ | |
|---|---|---|---|---|---|---|
| | Mean | $\sigma$ | Mean | $\sigma$ | Mean | $\sigma$ |
| Haar_1 | 0.5783 | 0.0763 | 0.1554 | 0.0488 | 0.7760 | 0.0455 |
| Spline_1 | 0.5618 | 0.0749 | 0.1546 | 0.0491 | 0.7483 | 0.0590 |
| Spline_2 | 0.5799 | 0.0783 | 0.1546 | 0.0475 | 0.7745 | 0.0467 |
| Spline_3 | **0.5857** | **0.0754** | 0.1568 | 0.0499 | **0.7767** | **0.0466** |
| LeGall 5/3 | 0.5769 | 0.0726 | 0.1569 | 0.0458 | 0.7743 | 0.0458 |
| CDF 9/7 | 0.5707 | 0.0723 | 0.1546 | 0.0521 | 0.7709 | 0.0468 |
| Rod 6/6 | 0.5775 | 0.0728 | **0.1570** | **0.0530** | 0.7741 | 0.0463 |

(b)

| Filter Bank | $Q_{AB/F}$ | | MI | | $Q_P$ | |
|---|---|---|---|---|---|---|
| | Mean | $\sigma$ | Mean | $\sigma$ | Mean | $\sigma$ |
| Haar_1 | 0.5949 | 0.0739 | **0.1574** | **0.0512** | **0.7751** | **0.0468** |
| Spline_1 | 0.5818 | 0.0739 | 0.1555 | 0.0502 | 0.7611 | 0.0530 |
| Spline_2 | 0.5934 | 0.0765 | 0.1566 | 0.0491 | 0.7727 | 0.0474 |
| Spline_3 | **0.5953** | **0.0736** | 0.1572 | 0.0509 | 0.7739 | 0.0476 |
| LeGall 5/3 | 0.5880 | 0.0694 | 0.1564 | 0.0528 | 0.7737 | 0.0465 |
| CDF 9/7 | 0.5788 | 0.0688 | 0.1533 | 0.0513 | 0.7672 | 0.0490 |
| Rod 6/6 | 0.5848 | 0.0701 | 0.1563 | 0.0525 | 0.7711 | 0.0478 |

(c)

**Table 4.2:** *Fusion results for IR-visible image pairs. (a) DTCWT and NSCT. (b) UWT without spectral factorization. (c) UWT with spectral factorization.*

for multifocus images, the application of filters with small support size yields no benefits, and, as can be seen in Table 4.4, best results are achieved using the NSCT and the DTCWT.

For IR-visible and medical image pairs the situation is substantially different. Since these image types come from different sensors, they exhibit a high degree of dissimilarity between different spectral bands. Hence, the application of filters with small support prior to the fusion process considerably improves the fusion result. We start our discussion by looking at Table 4.2, which lists the results for IR-visible image fusion. By looking at the second column, exhibiting the average results for the $Q_{AB/F}$ fusion metric, it can be noted that the proposed method yields

| Transform | $Q_{AB/F}$ | | MI | | $Q_P$ | |
|---|---|---|---|---|---|---|
| | Mean | $\sigma$ | Mean | $\sigma$ | Mean | $\sigma$ |
| DTCWT | 0.6314 | 0.0423 | 0.3853 | 0.0633 | 0.6618 | 0.0463 |
| NSCT | **0.6624** | **0.0415** | **0.4035** | **0.0556** | **0.6667** | **0.0397** |

(a)

| Filter Bank | $Q_{AB/F}$ | | MI | | $Q_P$ | |
|---|---|---|---|---|---|---|
| | Mean | $\sigma$ | Mean | $\sigma$ | Mean | $\sigma$ |
| Haar_1 | 0.6776 | 0.0464 | 0.4209 | 0.0452 | 0.6845 | 0.0309 |
| Spline_1 | 0.6507 | 0.0465 | 0.4191 | 0.0505 | 0.6594 | 0.0329 |
| Spline_2 | 0.6807 | 0.0469 | 0.4237 | 0.0509 | **0.6884** | **0.0286** |
| Spline_3 | **0.6834** | **0.0430** | **0.4248** | **0.0461** | 0.6852 | 0.0295 |
| LeGall 5/3 | 0.6614 | 0.0411 | 0.4035 | 0.0571 | 0.6631 | 0.0358 |
| CDF 9/7 | 0.6456 | 0.0422 | 0.3943 | 0.0618 | 0.6598 | 0.0392 |
| Rod 6/6 | 0.6641 | 0.0413 | 0.4069 | 0.0547 | 0.6630 | 0.0367 |

(b)

| Filter Bank | $Q_{AB/F}$ | | MI | | $Q_P$ | |
|---|---|---|---|---|---|---|
| | Mean | $\sigma$ | Mean | $\sigma$ | Mean | $\sigma$ |
| Haar_1 | **0.7100** | **0.0441** | 0.4289 | 0.0568 | 0.6687 | 0.0311 |
| Spline_1 | 0.6937 | 0.0431 | 0.4245 | 0.0583 | 0.6695 | 0.0382 |
| Spline_2 | 0.7094 | 0.0442 | **0.4313** | **0.0656** | **0.6719** | **0.0294** |
| Spline_3 | 0.7092 | 0.0432 | 0.4289 | 0.0571 | 0.6719 | 0.0328 |
| LeGall 5/3 | 0.6922 | 0.0420 | 0.4090 | 0.0583 | 0.6639 | 0.0375 |
| CDF 9/7 | 0.6827 | 0.0432 | 0.4000 | 0.0626 | 0.6631 | 0.0432 |
| Rod 6/6 | 0.6944 | 0.0444 | 0.4112 | 0.0577 | 0.6657 | 0.0397 |

(c)

**Table 4.3:** *Fusion results for medical image pairs. (a) DTCWT and NSCT. (b) UWT without spectral factorization. (c) UWT with spectral factorization.*

significantly better results for all filter banks under test, compared to the results for the DTCWT, NSCT and UWT without spectral factorization, suggesting that edges are better preserved using the UWT with spectral factorization. This is a particularly important result since the preservation of salient information is one of the main motivations of this work. In the case of the MI fusion metric, improvements are achieved for all non-orthogonal filter banks. On the other hand, for the $Q_P$ fusion metric the proposed methods yields no gains. Furthermore, it can be seen that the best scores are obtained for the non-orthogonal filter banks introduced in Section 4.3. Thus, this indicates that the increased filter design freedom of the UWT leads to filter banks which perform well in the context of IR-visible image

| Transform | $Q_{AB/F}$ | | MI | | $Q_P$ | |
|---|---|---|---|---|---|---|
| | Mean | $\sigma$ | Mean | $\sigma$ | Mean | $\sigma$ |
| DTCWT | 0.7327 | 0.0552 | **0.5104** | **0.0722** | 0.9070 | 0.0175 |
| NSCT | **0.7360** | **0.0552** | 0.5091 | 0.0722 | **0.9075** | **0.0179** |

(a)

| Transform | $Q_{AB/F}$ | | MI | | $Q_P$ | |
|---|---|---|---|---|---|---|
| | Mean | $\sigma$ | Mean | $\sigma$ | Mean | $\sigma$ |
| Haar_1 | 0.7219 | 0.0563 | 0.4846 | 0.0686 | 0.9039 | 0.0201 |
| Spline_1 | 0.7169 | 0.0624 | **0.5058** | **0.0733** | 0.8969 | 0.0248 |
| Spline_2 | 0.7209 | 0.0561 | 0.4901 | 0.0703 | 0.9041 | 0.0205 |
| Spline_3 | 0.7278 | 0.0571 | 0.5035 | 0.0725 | 0.9058 | 0.0205 |
| LeGall 5/3 | 0.7296 | 0.0567 | 0.5020 | 0.0741 | 0.9065 | 0.0201 |
| CDF 9/7 | 0.7307 | 0.0571 | 0.5054 | 0.0784 | 0.9065 | 0.0198 |
| Rod 6/6 | **0.7322** | **0.0555** | 0.5044 | 0.0747 | **0.9072** | **0.0191** |

(b)

| Transform | $Q_{AB/F}$ | | MI | | $Q_P$ | |
|---|---|---|---|---|---|---|
| | Mean | $\sigma$ | Mean | $\sigma$ | Mean | $\sigma$ |
| Haar_1 | 0.7315 | 0.0527 | 0.4869 | 0.0642 | 0.9037 | 0.0184 |
| Spline_1 | 0.7214 | 0.0563 | 0.4934 | 0.0715 | 0.8988 | 0.0231 |
| Spline_2 | 0.7294 | 0.0521 | 0.4854 | 0.0712 | **0.9041** | **0.0198** |
| Spline_3 | 0.7307 | 0.0531 | 0.4914 | 0.0697 | 0.9039 | 0.0200 |
| LeGall 5/3 | **0.7323** | **0.0527** | 0.4963 | 0.0711 | 0.9034 | 0.0198 |
| CDF 9/7 | 0.7262 | 0.0557 | 0.4963 | 0.0691 | 0.9000 | 0.0211 |
| Rod 6/6 | 0.7297 | 0.0542 | **0.4970** | **0.0694** | 0.9016 | 0.0208 |

(c)

**Table 4.4:** *Fusion results for multifocus image pairs. (a) DTCWT and NSCT. (b) UWT without spectral factorization. (c) UWT with spectral factorization.*

fusion. Finally, we would like to point out that the proposed spectral factorization framework significantly outperforms the fusion results obtained by state-of-the-art transforms such as the DTCWT and NSCT for all three fusion metrics.

The results of the fusion of an IR-visible image pair using the DTCWT, NSCT and UWT with and without spectral factorization are shown in Fig. 4.9. The 'Haar_-1' filter bank was employed in the UWT approaches. Examining the results on the zoomed images, illustrated in Figs. 4.9(e)-(h), the contours of the UWT-based fusion approaches seem to be slightly more accentuated. This is particularly visible when observing the persons' lower body part, displayed in the center of the image.

When examining the results shown in Table 4.3, the same conclusions can be

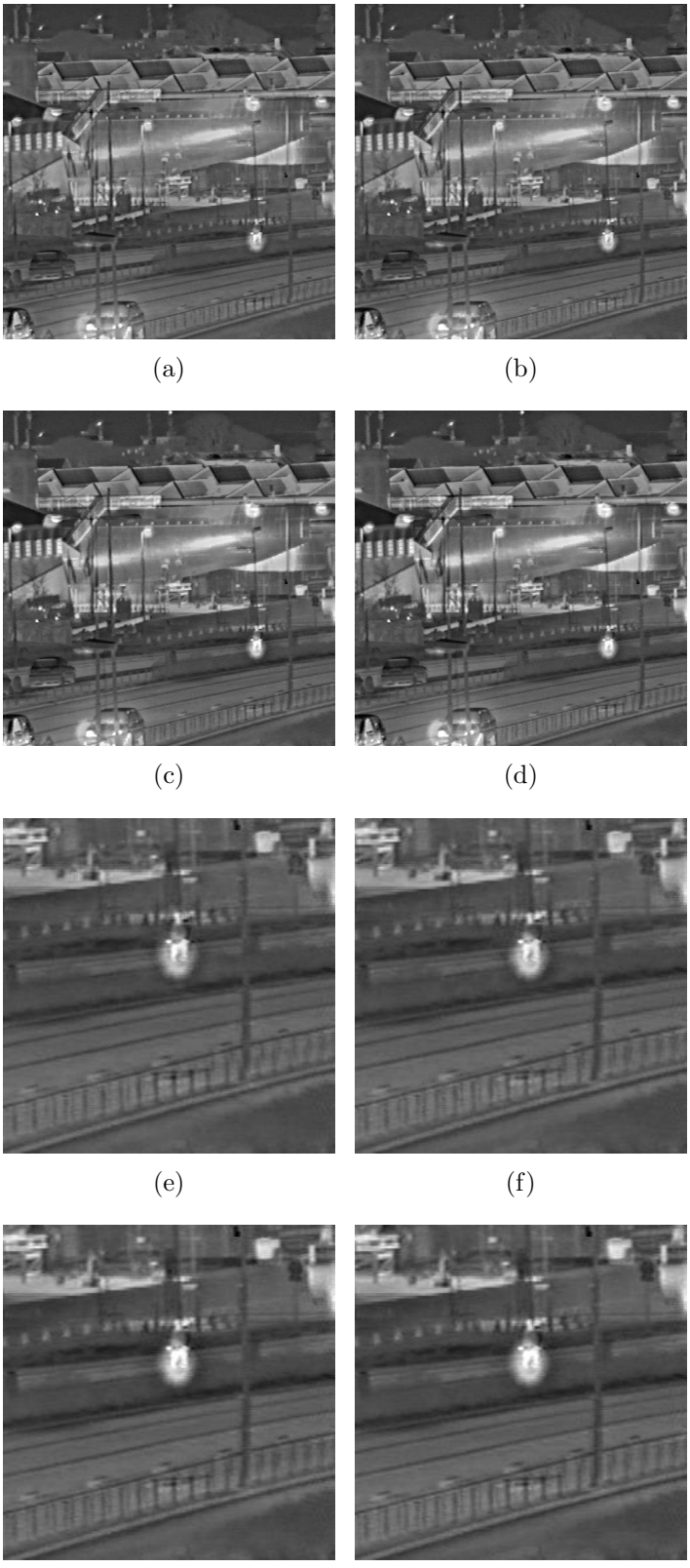


**Figure 4.9:** *Fusion results for an IR-visible image pair. (a) DTCWT fused. (b) NSCT fused. (c) UWT fused without spectral factorization. (d) UWT fused with spectral factorization. (e)-(h) Zoomed versions of (a)-(d).*

drawn for the set of medical images. However, since medical image pairs present, in general, an elevated number of regions, exhibiting information at coincident localizations, our approach yields even better results for these images than for the set of IR-visible images. This gain in fusion performance is most apparent when looking at the $Q_{AB/F}$ fusion score of the two image groups. Whereas for both image classes a considerable improvement is achieved for all filter banks, the gain is more than twice as high for medical image pairs. A similar tendency can be observed for the MI fusion metric, where the UWT fusion with spectral factorization produces a higher score for all tested filter banks, again suggesting the superiority of the proposed approach. In contrast, a moderate drop in fusion performance occurs for the $Q_P$ metric. However, it should be pointed out that this does not agree with subjective perception, as shown in the medical fusion example (Fig. 4.10). As before, best results are obtained when using the non-orthogonal filter banks of Section 4.3. Furthermore, the proposed method yields superior results for all three objective metrics when compared to conventional methods based on the NSCT and the DTCWT.

Fig. 4.10 shows the results for the fusion of a medical image pair, obtained by applying the DTCWT- and NSCT-based fusion scheme, as well as the UWT-based fusion scheme with and without spectral factorization in combination with the 'Haar_1' filter bank. Looking at the results obtained for the DTCWT and the NSCT, it can be observed that both schemes suffer from a significant loss of edge information, particularly noticeable at the outermost borders of the zoomed images (Figs. 4.10(e)-(h)). There, information belonging to the skull bone (white stripe enclosed within the gray, tube-like structure) partially disappeared. This is due to the superposition of the skull bones, originating from the medical source image pair, resulting in coefficient overlaps in the DTCWT and NSCT transform domain, which cannot be resolved by the fusion algorithm. As for the fusion results obtained with the UWT, this effect is reduced to a minimum and the edge information is preserved to a much higher degree. Moreover, in case of the UWT with spectral factorization, the edges appear to be more accentuated than in the fusion scenario without spectral factorization, thus indicating the perceptual superiority of the proposed spectral factorization approach.

To demonstrate the independence of the achieved results with respect to the underlying fusion rule, Figs. 4.11 and 4.12 show the average results for all IR-visible and medical image pairs, respectively, employing several different combination schemes. In more detail, we utilized the four fusion schemes discussed in Section 4.4 in combination with the DTCWT and NSCT, as well as with the UWT with and without our proposed spectral factorization approach (in Figs. 4.11 and 4.12 referred to as UWT and UWT-SF, respectively) and grouped the results in accordance with the used fusion metric. Table 4.5 gives an overview on the used fusion rules for all

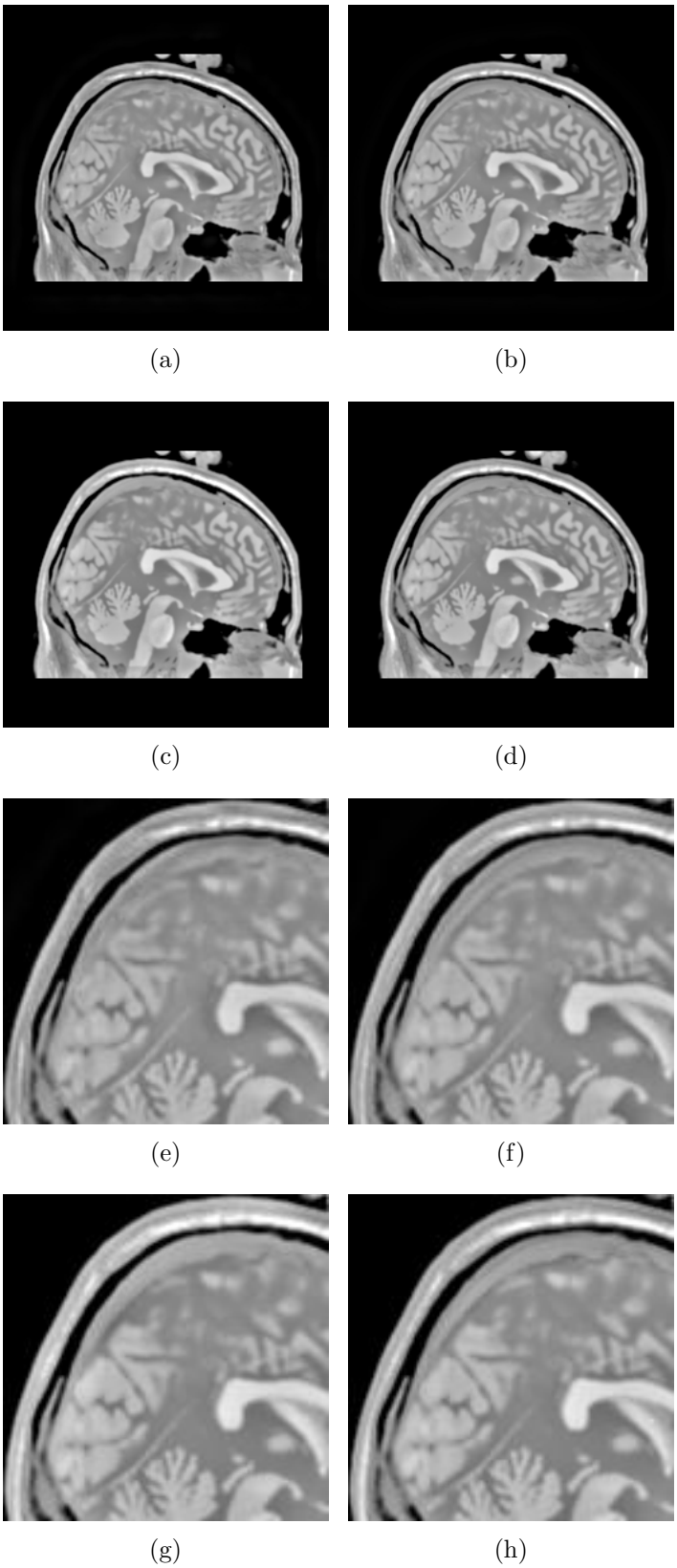
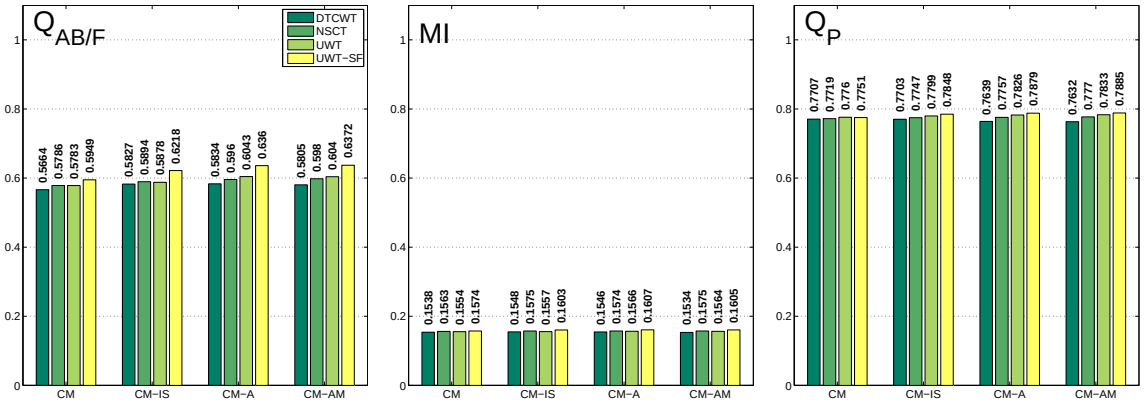
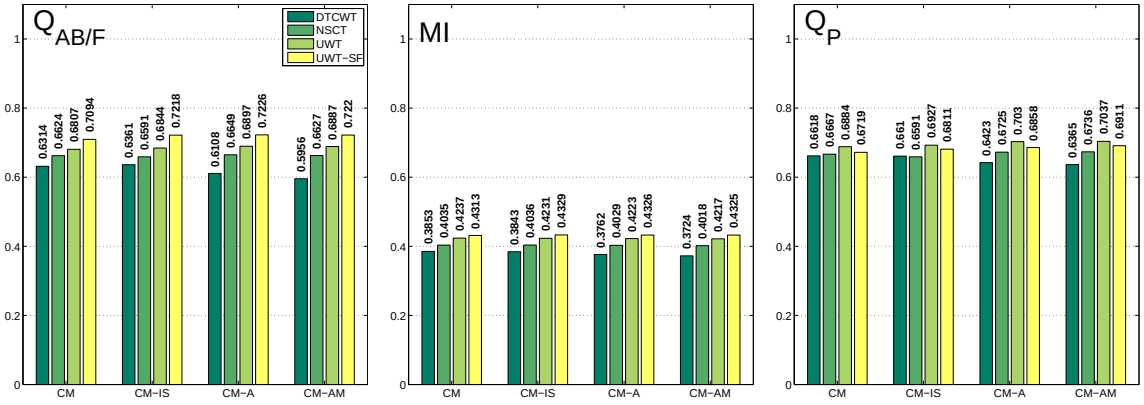


**Figure 4.10:** *Fusion results for a medical image pair. (a) DTCWT fused. (b) NSCT fused. (c) UWT fused without spectral factorization. (d) UWT fused with spectral factorization. (e)-(h) Zoomed versions of (a)-(d).*




(a) (b) (c)

**Figure 4.11:** *Comparison of different fusion rules for IR-visible image pairs using the (a) $Q_{AB/F}$, (b) MI and (c) $Q_P$ fusion metrics.*




(a) (b) (c)

**Figure 4.12:** *Comparison of different fusion rules for medical image pairs using the (a) $Q_{AB/F}$, (b) MI and (c) $Q_P$ fusion metrics.*

| Abbreviation | Description | Equation(s) |
|---|---|---|
| **CM** | *'Choose Max' fusion rule* | (3.51) |
| **CM-IS** | *CM with intra-scale grouping* | (4.10) |
| **CM-A** | *CM-IS with window-based activity measure* | (2.3), (4.11) |
| **CM-AM** | *Fusion rule by Burt and Kolczynski [32]* | (2.3)−(2.5), (4.12) |

**Table 4.5:** *Overview on the used fusion rules.*

detail images. The approximation images were fused using the averaging operation of eq. (3.52). As for the UWT-based approaches, the 'Haar_1' filter bank was employed for all IR-visible image pairs whereas the 'Spline_2' filter bank was used for the set of medical image pairs. By observing the results it can be noted that for all investigated fusion schemes the best results are achieved using the proposed spectral factorization method. In fact for IR-visible image pairs it only ranks second for the $Q_P$ fusion metric together with the CM fusion rule, whereas for medical image pairs

it gains first place for the $Q_{AB/F}$ and MI fusion metric and only ranks second for the $Q_P$ score. Note that this is in accordance with the results presented in Tables 4.2 and 4.3. Two important conclusions can be drawn from this observation: a) The introduced fusion framework with spectral factorization indeed tends to generate the best multiscale fusion results independent of the employed fusion rule and b) no tested combination scheme was able to resolve the problems originating from the superposition of coefficient values within the same spectral band. Consequently, since the probability of coefficients with coincident localizations can be directly associated with the support length of the applied filter bank, our proposed framework with spectral factorization can in fact be considered as a good alternative to alleviate the original problem.

## 4.6 Conclusions

A novel UWT-based pixel-level image fusion approach is presented in this chapter. It successfully improves fusion results for images exhibiting features at nearby located and coincident pixel locations. Our method spectrally divides the analysis filter pair into two factors which are then separately applied to the input image pair, splitting the image decomposition procedure into two successive filter operations. The actual fusion step takes place after convolution with the first filter pair. It is equivalent, as far as the coefficient spread is concerned, to a filter with significantly smaller support size than the original filter pair. Thus, the effect of the coefficient spreading problem, which tends to considerably complicate the feature selection process, is successfully reduced. This leads to a better conservation of features which are located close to each other in the input images. In addition, this solution leaves room for further improvements by taking advantage of the nonsubsampled nature of the UWT, which permits the design of non-orthogonal filter banks where both synthesis filters exhibit only positive coefficients. Such filters provide a reconstructed, fused image less vulnerable to ringing artifacts.

The obtained experimental results have been analyzed in terms of the three objective metrics $Q_{AB/F}$, MI and $Q_P$. They showed that for multisensor images, such as IR-visible and medical image pairs, the proposed spectral factorization framework significantly outperforms fusion schemes based on state-of-the-art transforms such as the DTCWT and the NSCT, independent of the underlying fusion rule. Additionally, the perceptual superiority of the proposed framework was suggested by informal visual inspection of a fused IR-visible as well as a fused medical image pair.

# Chapter 5

# Infrared-visible image fusion using the Undecimated Wavelet Transform with spectral factorization and target extraction

In this chapter we propose an extension to the fusion framework of Chapter 4 by including information about the presence of targets within the infrared (IR) image to the fusion process. For this purpose we introduce a novel IR segmentation method which is able to detect targets in low-contrast environments without introducing spurious results. Steered by the segmentation process we ensure that the most relevant information from the IR image is included in the fused image, leading to a more accurate representation of the captured scene. Since the target extraction is performed on the decomposed images obtained after application of the first spectral factor, it can be embedded directly within the existing fusion framework. Additionally, a new hybrid fusion scheme is proposed in this chapter which utilizes both pixel-level and region-level information to fuse the source images, turning the fusion process more robust against possible segmentation errors which may corrupt the final composite image. The combination of these techniques leads to a novel fusion framework which is able to improve the fusion results of its pure pixel-level counterpart without target extraction. Furthermore, traditional pixel-level fusion approaches, based on state-of-the-art transforms such as the Nonsubsampled Contourlet Transform (NSCT) and the Dual-Tree Complex Wavelet Transform (DTCWT), are significantly outperformed by the use of the proposed set of methods.

The structure of this chapter is as follows: Section 5.1 introduces the target extraction algorithm based on a marker-controlled watershed transformation. Its inclusion into the existing fusion framework of Chapter 4 is the topic of Section 5.2

whereas Section 5.3 discusses the obtained simulation results and compares them with other state-of-the-art fusion schemes. Finally, our conclusions are given in Section 5.4. We make use of the same notation, where suitable, as given in Section 2.1.

## 5.1 Target extraction algorithm

A number of segmentation techniques have been proposed for the purpose of image fusion, e.g. [7], [8], [10] and [47]. Most of these methods first employ a multiscale transform to the source images and extract the regions from the transform coefficients.

In general, the fusion performance of region-based image fusion methods highly depends on the quality of the segmentation process. For example, objects-of-interest which are concealed within other regions may not be incorporated in the fused image. On the other hand, features which are split into more than one region may cause unwanted side effects such as ringing effects in the fused image [47]. Unfortunately, in case of IR-visible image fusion, a proper segmentation map for all input images is difficult to achieve due to the different nature of the imaging sensors.

The approach taken in this chapter varies substantially from conventional region-based fusion approaches. The main difference is that we do not segment both IR and visible images with the help of a single segmentation algorithm. Instead we use a priori knowledge of the properties of IR images to successfully extract all objects-of-interest. An IR image is the result of the acquisition of thermal radiation of a scene, producing a 2-D map representing the temperature, emissivity and reflexivity variation of the respective scene [115]. Thus, we can define an object-of-interest (or target) as an enclosed region with either a larger or smaller temperature than the environment which is situated beyond transient regions such as edges.

In this chapter we propose the use of a marker-controlled watershed transformation to extract possible targets from the IR image. The marker image is computed using the gradient modulus maxima of the Undecimated Wavelet Transform (UWT) in combination with a novel edge tracking approach. The block diagram of our proposed target extraction method is given in Fig. 5.1. It can be considered as consisting of three main parts: marker extraction, image simplification and watershed transformation. In the remainder of this section, these steps are explained in detail. Note that all employed thresholds were determined empirically from the set of available IR images (see Fig. 5.11).

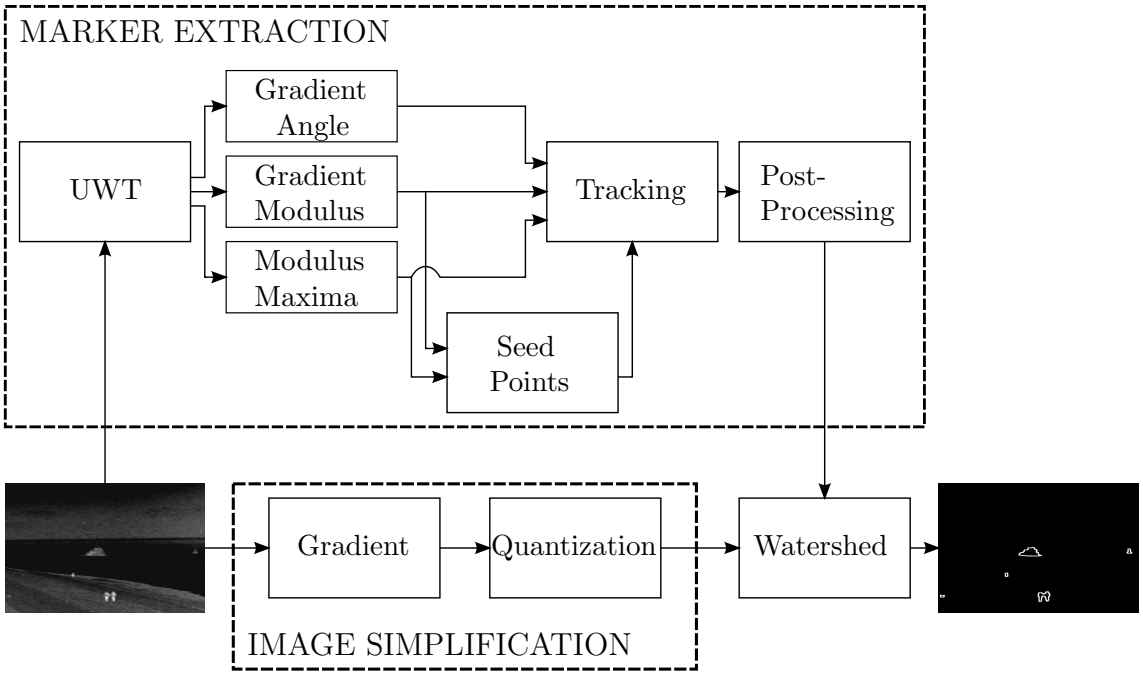



**Figure 5.1:** *Block diagram of the proposed target extraction approach.*

### 5.1.1 Marker extraction

The direct application of the watershed transformation usually leads to a considerable over-segmentation of the input image. One way to improve the results is the use of the watershed transformation in combination with a marker image, limiting the segmentation process to some "marked" areas [116]. Since targets in IR images are usually bounded by transient regions such as edges, it seems natural to use some sort of edge detector for this task. In this work we propose the use of the UWT-based multiscale edge detector of [117], which was modified so that it can be seamlessly integrated in the pre-existing fusion framework of Chapter 4.

In our approach, the input image is first decomposed in a set of approximation images $x^j$ at different scales $j$ by separate convolution of the rows and columns with an upsampled low-pass filter $h^{(j)}$ as described in eq. (3.26). Next, we calculate the horizontal and vertical detail images $y_1^{j+1}$ and $y_2^{j+1}$ for each scale by employing the upsampled first spectral factors $(1 + z^{-2^{j-1}})$ and $(1 - z^{-2^{j-1}})$ of eq. (4.2) to the approximation images $x^j$. Note that this corresponds to the filtering with an upsampled Haar filter pair in the spatial domain. Due to the nature of the Haar filter, the resulting detail images can be considered as the directional derivatives of the approximation images. Thus, we can define the gradient vector $\vec{\nabla} x^j$ at each position $m, n$ and scale $j$ as consisting of the set of horizontal and vertical detail

images such that

$$\begin{pmatrix} y_1^{j+1}[m,n] \\ y_2^{j+1}[m,n] \end{pmatrix} = \vec{\nabla}\left(x^0 * h^{(j)}\right)[m,n] = \vec{\nabla}x^j[m,n], \tag{5.1}$$

where $x^0$ represents the input IR image.

Based on the gradient vector, three images are calculated at each scale which will subsequently be used in the tracking step (see Fig. 5.1). These are the modulus of the gradient vector $\vec{\nabla}x^j$ given by

$$\left|\vec{\nabla}x^j[m,n]\right| = \sqrt{y_1^{j+1}[m,n]^2 + y_2^{j+1}[m,n]^2}, \tag{5.2}$$

the angle of the steepest ascent of the gradient vector

$$\measuredangle\vec{\nabla}x^j[m,n] = \arctan\left(\frac{y_2^{j+1}[m,n]}{y_1^{j+1}[m,n]}\right), \tag{5.3}$$

and a binary image containing the positions of the local modulus maxima of the gradient vector, corresponding to the zero-crossings of the second-order directional derivatives of $x^j$. Figs. 5.2(b)-(d) show the three resulting images at the 3$^{\text{rd}}$ decomposition level for the sample IR image of Fig. 5.2(a).

As a next step, the binary gradient modulus maxima images are multiplied with the corresponding gradient modulus images and a first threshold is applied. This results in a binary image containing only those gradient modulus maxima above the chosen threshold. After combining the thresholded images of the 1$^{\text{st}}$ and 2$^{\text{nd}}$ decomposition level using a logical AND operation, we obtain a first coarse segmentation as depicted in Fig. 5.2(e). From this preliminary segmentation, the seed/starting points for the subsequent edge tracking operation are computed by extracting the endpoints of the segmented stubs.

The proposed tracking algorithm starts by taking a seed point from the seed point list and follows the target border in the direction perpendicular to the gradient angle, marking each encountered pixel on its way as belonging to a target edge. Thus, in order to track a target it is sufficient that a single seed point is located on the target edge. This permits the selection of a high initial threshold, minimizing the introduction of false targets in the segmentation process. At each new point, the tracking algorithm multiplies the 8-connected neighborhood of the tracked pixel with a directional mask, discarding those pixels which do not agree with the mask's angle. Note that in order to turn the tracking direction more robust against possible angular outliers, the utilized angle is computed by averaging the gradient angles of all but the first decomposition level. The directional masks with their corresponding

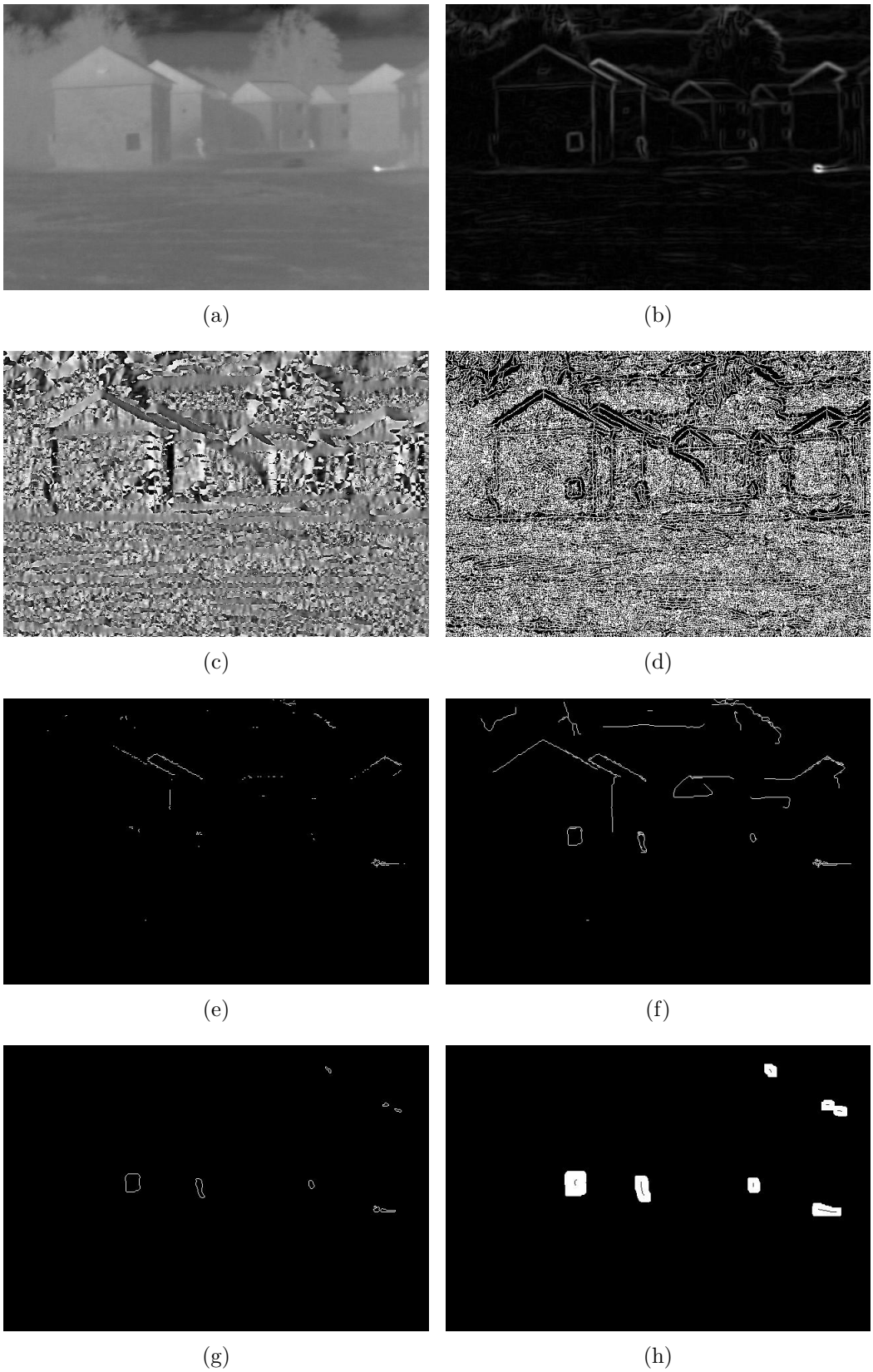


**Figure 5.2:** *Results of the marker extraction. (a) Original IR image. (b) Gradient modulus image ($3^{rd}$ decomposition level). (c) Gradient angle image ($3^{rd}$ decomposition level). (d) Gradient modulus maxima image ($3^{rd}$ decomposition level). (e) Preliminary segmentation map used for seed point extraction. (f) Tracked image. (g) Post-processed, tracked image. (h) Binary marker image.*

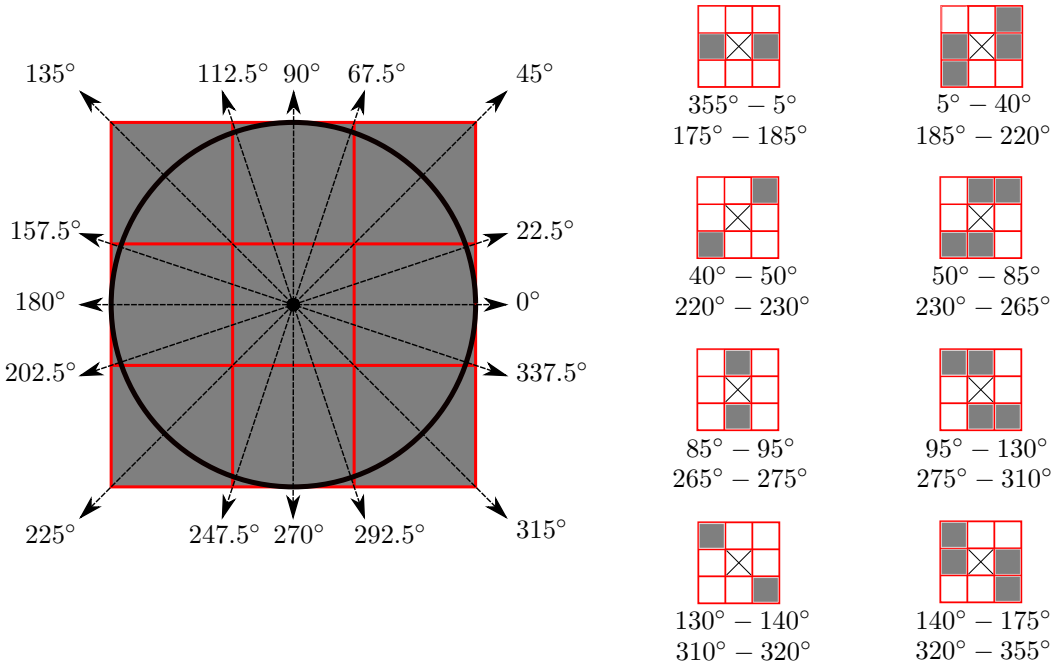
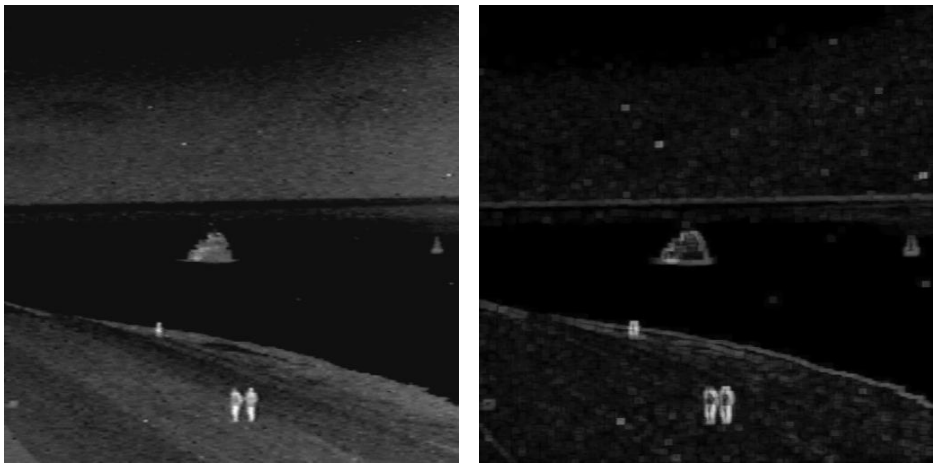



**Figure 5.3:** *Directional masks of the tracking operation.*

directions are given in Fig. 5.3.

From all candidate pixels (gray pixels depicted in the directional masks on the right-hand side of Fig. 5.3), the tracking algorithm chooses the one with the highest average gradient modulus and labels it as tracked. Additionally, the 4-connected neighbors, as well as all remaining candidate pixels arising from the previously tracked pixel, are marked as "discarded", avoiding the use of these pixels as candidate pixels again. Note that this step is of prime importance since it circumvents the ambiguity problem of the gradient angle (there always exist two tracking paths pointing in opposite directions). The tracking stops if: 1) the new, tracked point is 8-connected to a previously tracked point or 2) the averaged gradient modulus is below a empirically determined threshold. Fig. 5.2(f) shows the result of the tracking operation.

Following our definition of a valid target (see definition above) an object-of-interest always forms a bounded region. Thus, we apply a post-processing step which cleans the tracked image by removing all edge-segments which do not form a closed region. Furthermore, in order to make the result more robust against spurious targets, we remove all objects smaller than 40 pixels from the tracked image. The post-processed, tracked image is illustrated in Fig. 5.2(g). It can be observed that all objects-of-interest, originating from the IR image of Fig. 5.2(a), are successfully conserved. Note that the tracked image of Fig. 5.2(g) also exhibits some spurious targets. However, since these wrongly tracked regions do not correspond to any bounded object in the source image, they form regions of very small size after application of the watershed transformation and can thus be removed easily



(a) (b)

**Figure 5.4:** *Result of the image simplification process. (a) Original IR image. (b) Simplified IR image after application of the morphological gradient followed by quantization.*

thereafter.

In order to ensure that the entire target is included in the marker image, the target area is filled and dilated using a $6 \times 6$ square structure element. Moreover, the watershed process demands that each marked region exhibits, at least, one pixel at the marker center which does not belong to the marked target area (black pixels enclosed by the white, marked regions in Figs. 5.2(h) and 5.7(b)). This is achieved by performing a skeletonization of the filled, dilated image followed by combining its outcome with the filled and dilated image using a logical XOR operation. The final binary marker image is given in Fig. 5.2(h).

### 5.1.2 Image simplification

Before performing the marker-controlled watershed transformation it is advantageous to simplify the original IR image [116]. The approach adopted in this work is similar to the one proposed in [115]. More specifically, we simplify the source IR image by computing the morphological gradient, defined as the arithmetic difference between a dilation and an erosion, using a $6 \times 6$ structuring element, followed by a quantization of the resulting gradient image to 100 gray levels. The result of the simplification process, employed to the source IR image in Fig. 5.4(a), is illustrated in Fig. 5.4(b).

### 5.1.3 Watershed Transformation

The use of the morphological watershed transformation has been proven to be a powerful technique for segmenting images in many situations. It was first mentioned in a work by Beucher and Lantuejoul [118] who used the concept of watersheds (or

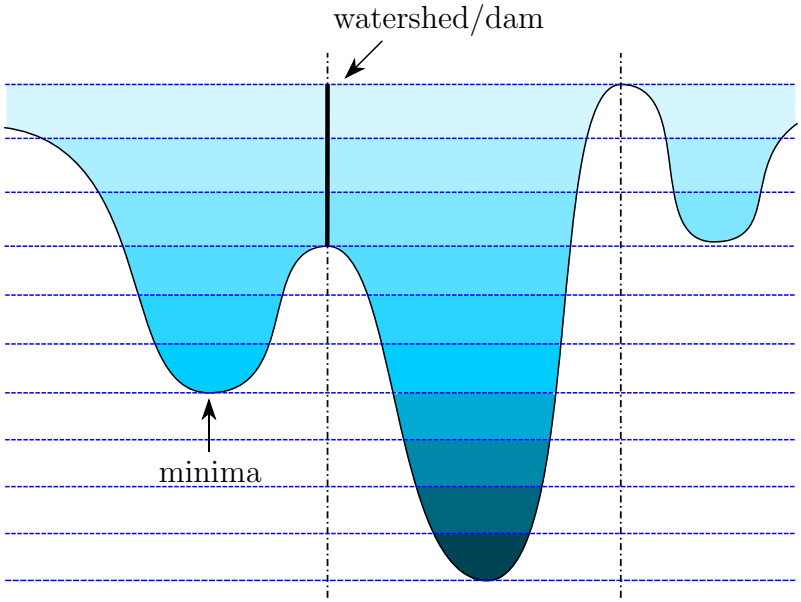
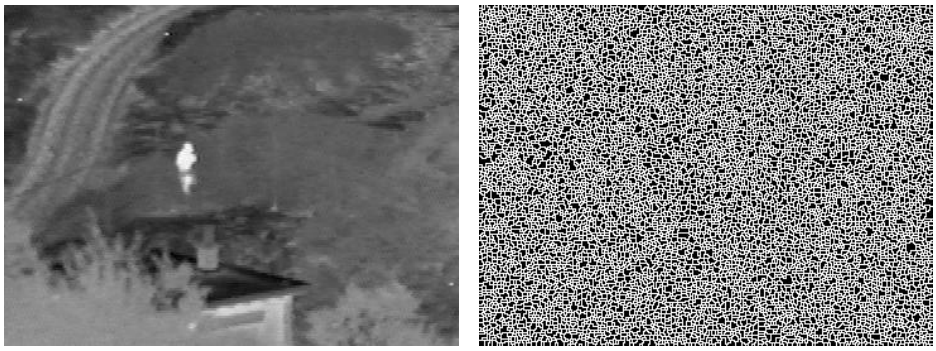



**Figure 5.5:** *Schematic illustration of the Watershed Transformation, according to the flooding scheme.*

dams) for bubble detection in radiographic plates and facet detection in fractures in steel. In what follows we briefly introduce the main idea behind the watershed transformation, based on the so-called flooding scheme as presented in [116], before describing its integration in the proposed target extraction framework. Please note that henceforth we consider a gray-level image as a topographic surface, where a light gray tone of a pixel corresponds to a high altitude on the topographic surface.

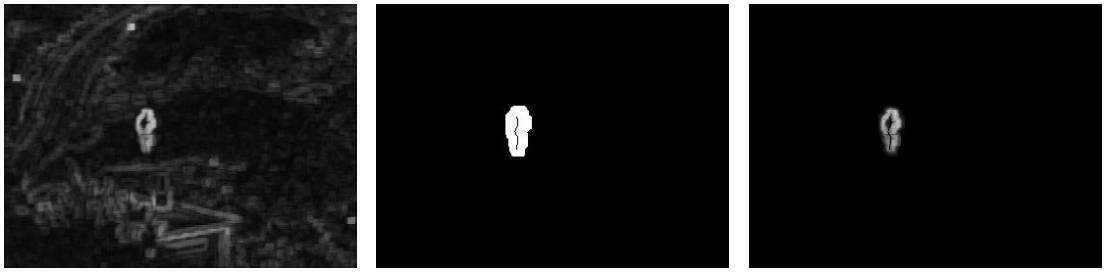
In order to perform the watershed transformation one usually starts by calculating the modulus of the gradient of the input image which may be obtained by assigning to each pixel $m, n$ the difference between the highest and the lowest pixels within a given neighborhood of $m, n$. In the corresponding topographic surface of the gradient modulus image, the highest values belong to regions with high contrast in the original image. Furthermore, each local minimum/maximum in the original image becomes a local minimum (valley) in the gradient modulus image surrounded by a closed chain of mountains, like a basin. The concept of the watershed transformation is now as follows: Imagine we bore a hole in each minimum of the topographic surface of the gradient modulus image and immerse it in a lake. The water entering through these holes fills up the various catchment basins. Now, in order to avoid the confluence of the floods coming from different minima, we build dams along the lines where the floods would merge. After complete immersion only the dams emerge and separate the various catchment basins, representing the outcome of the segmentation process. Fig. 5.5 schematically illustrates this process.

As already elaborated on previously, the direct application of the watershed transformation tends to yield a severe over-segmentation. This is mainly due to the high sensitivity of the gradient image to noise, leading to many negligible re-



(a) (b)

**Figure 5.6:** *Over-segmentation caused by the Watershed Transformation. (a) Original IR image. (b) Result of the watershed tranformation when applied directly to (a).*



(a) (b) (c)

**Figure 5.7:** *Construction of the input image for the watershed transformation. (a) Simplified IR image. (b) Binary marker image. (c) Input image of the watershed transformation (pixel-wise minimum of (a) and (b)).*

gional minima as illustrated in Fig. 5.6. In the proposed methodology, this over-segmentation is avoided by a) the use of a marker image which restricts the segmentation process to some highlighted regions-of-interest and b) a simplification of the input IR image, reducing the number of insignificant regional minima. More specifically, the watershed transformation is employed to the image obtained by calculating the pixel-wise minimum between the binary marker image of Section 5.1.1 and the simplified IR image of Section 5.1.2. Fig. 5.7 visualizes this process for the IR source image of Fig. 5.6(a). Note that after application of the watershed transformation all objects which do not exceed an overall size of 40 pixels are again removed from the segmented image. The final result of the target extraction stage for the three IR images depicted in Figs. 5.2(a), 5.4(a) and 5.6(a) can be seen in Fig. 5.8.

In the next section we demonstrate how the extracted target information is utilized during the fusion process.

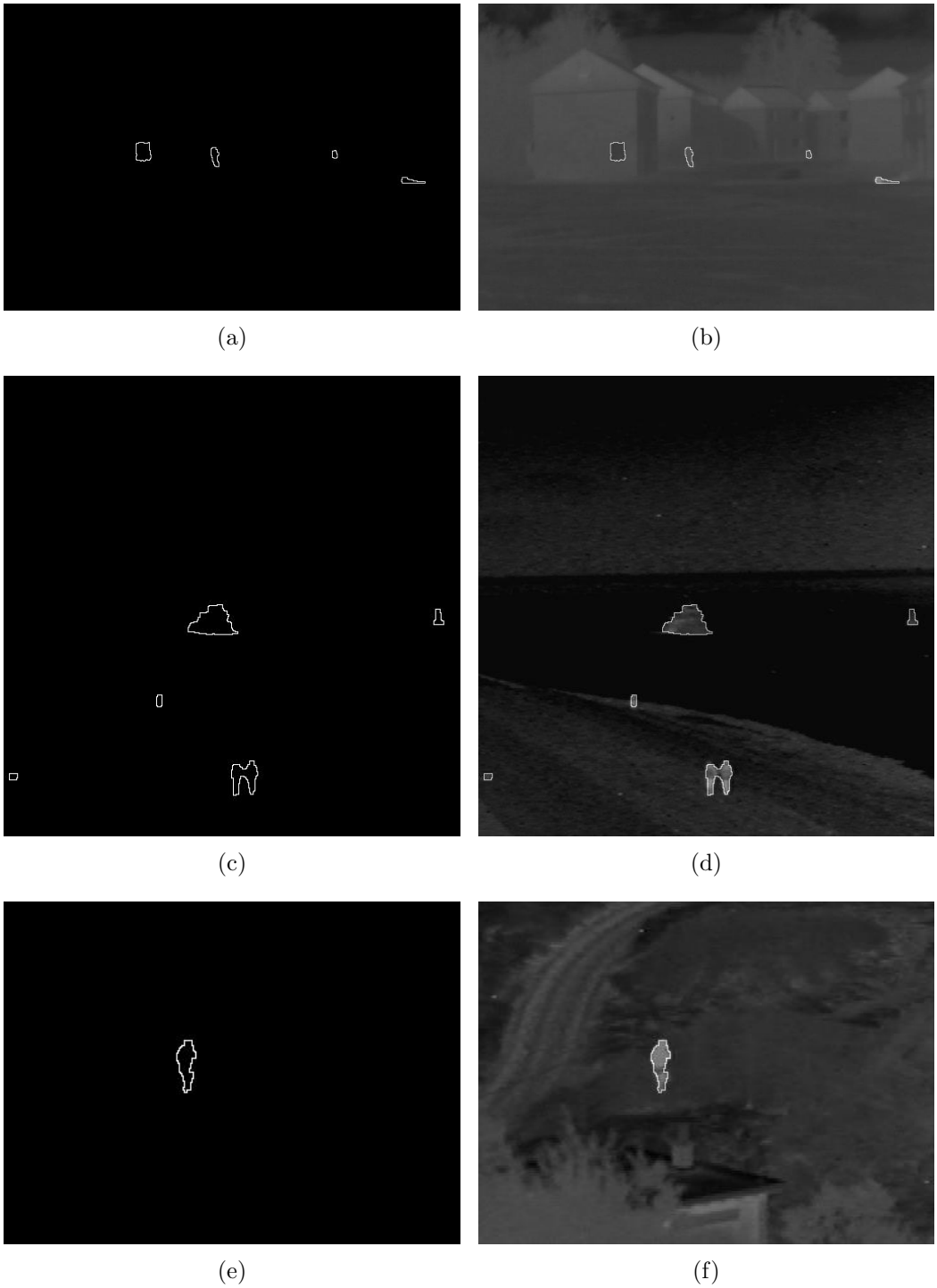


**Figure 5.8:** *Results of the target extraction. (a), (c), (e) Binary segmentation maps. (b), (d), (f) Binary segmentation maps superimposed on the corresponding IR source images of Figs. 5.2(a), 5.4(a) and 5.6(a).*

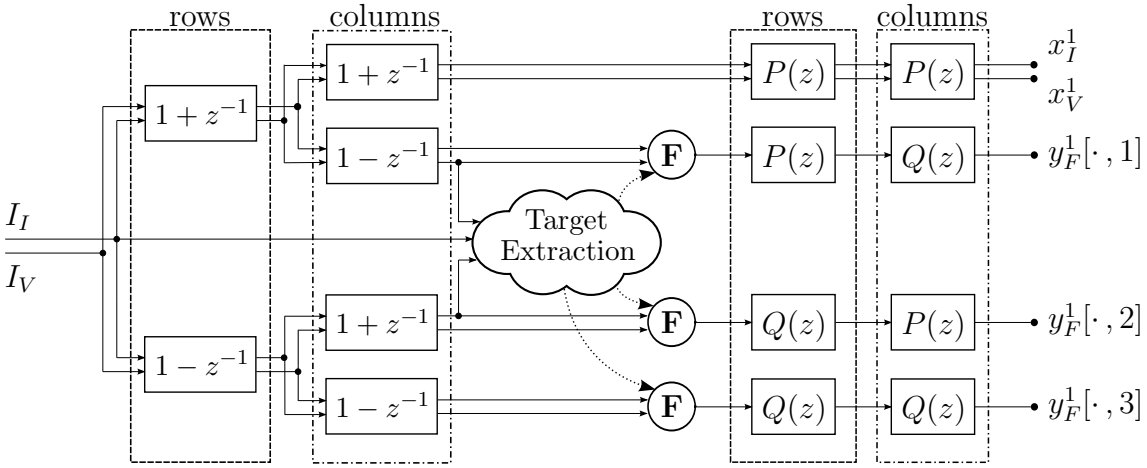



**Figure 5.9:** *Implementation of the $1^{st}$ stage of the proposed hybrid fusion framework.*

## 5.2 Overall fusion framework

In most region-level fusion methods, the actual fusion process is solely concerned with the proper combination of the segmented regions. This is usually done by weighted averaging of associated regions within the source images. Even though this technique has been shown to be effective, its performance highly depends on the quality of the computed region map. In other words, segmentation errors such as under- or over-segmentation may lead to the absence or degradation of certain features in the fused image, respectively.

In this work, the use of a hybrid fusion scheme is proposed. Here, all extracted targets are fused using a region-level fusion rule whereas the remaining image portions are fused by employing the pixel-level fusion rules, given in eqs. (4.10) and (3.52). This turns the fusion process more robust against the introduction of segmentation-induced fusion errors since we can still rely on the pixel-level algorithm to correctly incorporate an object-of-interest in the fused image, in case it was "missed" by the segmentation process. Fig. 5.9 shows the implementation of the proposed, overall fusion framework, combining the UWT-based fusion approach with spectral factorization of Chapter 4 with the target extraction algorithm of Section 5.1, for the $1^{\text{st}}$ decomposition level. In this chapter we are solely concerned with the fusion of a single, registered IR-visible image pair. However, the presented fusion scheme can easily be extended to the case of multiple input images.

After decomposing the input images using the first spectral factors $(1+z^{-2^{j-1}})$ and $(1-z^{-2^{j-1}})$, respectively, we apply the target extraction algorithm to the horizontal and vertical detail images of the IR image. Subsequently, the extracted target information is used to guide the fusion process. In this context we differentiate between two fusion scenarios which are introduced next.

### 5.2.1 Fusion of non-target regions

The first scenario is concerned with the fusion of transform coefficients not belonging to any extracted target. In this case the following fusion rules are used: The detail coefficients $y_I^j[m,n,p]$ and $y_V^j[m,n,p]$ of the IR and visible image, respectively, are fused using the pixel-level "choose max" fusion rule with intra-scale grouping as stated in eq. (4.10). Thereby, we ensure that the fusion decision at each decomposition level $j$ and spatial location $m,n$ is taken jointly for all three orientation bands $p$. The approximation coefficients $x_I^J[m,n]$ and $x_V^J[m,n]$ at the coarsest decomposition level $J$ are combined using the simple averaging operation given in eq. (3.52).

### 5.2.2 Fusion of target regions

A different approach is adopted for all transform coefficients belonging to an extracted target region. First, a measure of the matching degree between the transform coefficients (belonging to a single target region) of the IR and visible image is calculated. Consequently, each extracted target is classified as being present only in the IR image or in both source images. Based on this classification the following semantic region-level fusion rule is derived: If the extracted target is not evident in the visible image (unambiguous target), all detail and approximation coefficients of the corresponding region are directly transferred from the IR decomposition to the fused decomposition. Otherwise, the fusion will be handled by the pixel-based fusion scheme as discussed in Section 5.2.1. Please note that we expand the extracted target region in each decomposition step by $(2^{j-1}-1)$ pixels in all directions. Thereby, we compensate for the coefficient spread occurring at each decomposition level due to the filtering involved.

In order to calculate the match metric between the same target regions within the IR and visible image, two metrics are considered.

#### Match metric by Piella

The first match metric measures the normalized correlation between the transform coefficients averaged over the target region $\mathcal{R}_k$ for each decomposition level $j$ and direction $p$ as given in [119]

$$M_1^j(\mathcal{R}_k,p) = \frac{2\sum\limits_{(m,n)\in\mathcal{R}_k} y_I^j[m,n,p]\,y_V^j[m,n,p]}{\sum\limits_{(m,n)\in\mathcal{R}_k} \left|y_I^j[m,n,p]\right|^2 + \left|y_V^j[m,n,p]\right|^2}. \tag{5.4}$$

The final match measure is obtained after taking the absolute value of the averaged metrics, thus, bounding the final result to the interval $[0,1]$ with a value close to

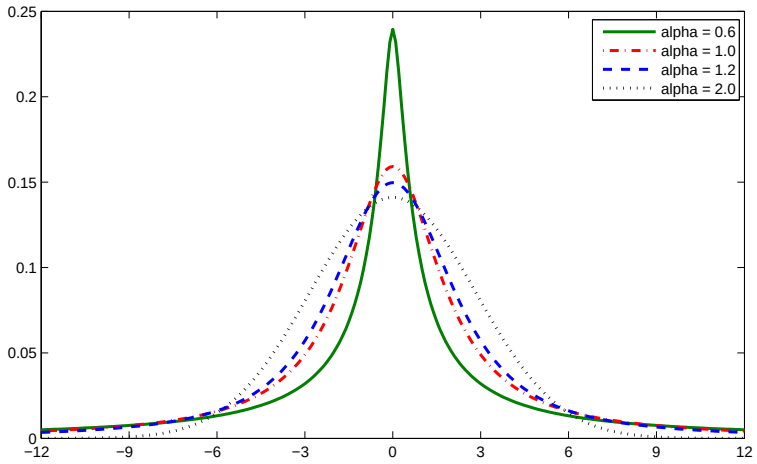



**Figure 5.10:** *Probability density functions of the SαS distribution corresponding to four different values of the characteristic exponent* $\alpha$*. The remaining parameters* $\gamma$ *and* $\delta$ *are fixed to* 2 *and* 0*, respectively.*

one suggesting a high similarity between the compared regions.

### Match metric based on alpha-stable modeling of wavelet coefficients

Alternatively, a second match metric is implemented which first models the wavelet coefficients of each target region as symmetric alpha-stable (S$\alpha$S) random processes. This choice is motivated by the fact that statistical distributions with heavy algebraic tails, such as the S$\alpha$S family, are considered to be accurate modeling tools for the wavelet coefficients of images [120]. Due to the lack of a compact analytical expression for the probability density function, S$\alpha$S distributions are best defined by their characteristic function [47]

$$\varphi(\omega) = \exp\left(j\delta\omega - \gamma|\omega|^{\alpha}\right), \tag{5.5}$$

where $\alpha$ is the characteristic exponent, $\delta$ is the location parameter, and $\gamma$ is the dispersion of the distribution. Fig. 5.10 shows the S$\alpha$S density functions for four different values of the characteristic exponent $\alpha$. It can be noticed that the smaller the characteristic exponent is, the heavier the tails of the S$\alpha$S probability density function. This implies that random variables following S$\alpha$S distributions with small characteristic exponents are highly impulsive [47].

By assuming that the location parameter $\delta$ is zero in the wavelet domain, we can estimate the two parameters $\alpha$ and $\gamma$ by calculating the first two logarithmic absolute moments of the wavelet coefficients, as described in [121]. More specifically, let us define $X$ as being a S$\alpha$S random variable, consisting of the set of wavelet coefficients at decomposition level $j$ and direction $p$ corresponding to an arbitrary

target region $\mathcal{R}_k$ and $Y$ as being the corresponding $\log|\mathrm{S}\alpha\mathrm{S}|$ random variable such that $Y = \log|X|$. Now it can be shown [121] that the mean and variance of $Y$ are related to the parameters $\alpha$ and $\gamma$ by

$$E(Y) = C_e\left(\frac{1}{\alpha} - 1\right) + \frac{1}{\alpha}\log\gamma \tag{5.6}$$

and

$$\mathrm{Var}(Y) = E([Y - E(Y)]^2) = \frac{\pi^2}{12}\frac{\alpha^2 + 2}{\alpha^2}, \tag{5.7}$$

where $C_e = 0.57721566\ldots$ is the Euler constant [122]. Thus, the estimation process involves solving eq. (5.7) for $\alpha$ and substituting back in eq. (5.6) to find the value of the dispersion parameter $\gamma$.

Next, the similarity between two corresponding target regions is calculated by means of the Kullback-Leibler distance (KLD). In information theory, the KLD or relative entropy is a measure of the distance between two distributions and is defined as [123]

$$D(p\|q) = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)}, \tag{5.8}$$

where $p(x)$ and $q(x)$ are two probability density functions (PDF). The KLD is always nonnegative and is zero if and only if $p = q$. However, it is not a true distance between distributions since it is not symmetric and does not satisfy the triangle inequality. Furthermore, following the convention that $0\log\frac{0}{q} = 0$ and $p\log\frac{p}{0} = \infty$, there may exist a symbol $x \in \mathcal{X}$ such that $p(x) > 0$ and $q(x) = 0$. This would yield that $D(p\|q) = \infty$, indicating that there may not always exist an upper bound for the KLD of two PDFs [123].

There exists no closed-form expression for the KLD between two general S$\alpha$S distributions. However, the KLD can be applied on the normalized versions of the corresponding characteristic functions [120]. In more detail, if we denote by $\alpha_1$, $\gamma_1$ and $\alpha_2$, $\gamma_2$ the extracted model parameters of target region $\mathcal{R}_k$ at decomposition level $j$ and direction $p$, derived from the IR and visible image, respectively, the KLD can be defined as [47]

$$M_2^j(\mathcal{R}_k, p) = \ln\left(\frac{c_2}{c_1}\right) - \frac{1}{\alpha_1} + \frac{2\gamma_2\Gamma\left(\frac{\alpha_2+1}{\alpha_1}\right)}{c_1\alpha_1\gamma_1^{\frac{\alpha_2+1}{\alpha_1}}} \tag{5.9}$$

with

$$c_i = \frac{2\Gamma\left(\frac{1}{\alpha_i}\right)}{\alpha_i\gamma_i^{1/\alpha_i}} \qquad i = 1, 2\ , \tag{5.10}$$

where $\Gamma(\cdot)$ represents the Gamma function. Note that in this case a value close to

**Figure 5.11:** *Thumbnails of all IR-visible image pairs used for evaluation purposes. Top row consists of IR images, whereas the bottom row represents the corresponding visible images.*

zero indicates a high resemblance between the two target regions. ■

The final classification is obtained after applying a threshold to the computed similarity scores, where in case of Piellas' match metric $M_1^j$ all targets below it, and in case of the S$\alpha$S model-based match metric $M_2^j$ all targets above it, are transferred directly to the fused decomposition. After the fusion step is complete, the filter pair represented by the $2^{\text{nd}}$ spectral factor ($P(z)$ and $Q(z)$ in eq. (4.1)) is applied to the approximation images and the fused detail images. Once the desired number of decompositions is reached, the approximation images are merged and the fused image is computed by applying the inverse UWT, using the corresponding synthesis filter bank without spectral factorization.

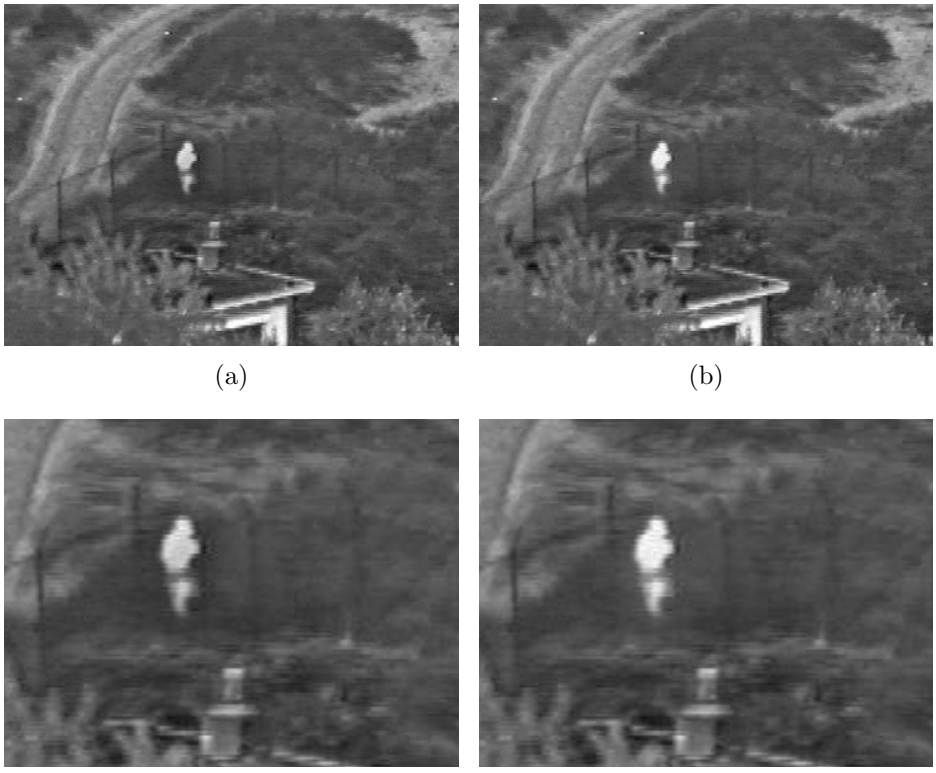
## 5.3 Results

The performance of the proposed image fusion scheme with target extraction was compared to the pixel-level fusion results obtained by applying the DTCWT, the NSCT and the UWT with spectral factorization (UWT-SF) of Chapter 4. As for the DTCWT and the NSCT we followed the same transform settings as listed in Table 4.1. In case of the UWT-based fusion schemes, we chose the non-orthogonal 'Haar_-1' filter bank of eq. (4.9). Please note that in this approach both synthesis filters $\tilde{h}$ and $\tilde{g}$ are positive and do not oscillate, hence providing a fused reconstruction less vulnerable to ringing artifacts. Four decomposition levels were chosen for all transforms.

We performed simulations for 5 IR-visible image pairs depicting a varying number of target regions within the IR image. The thumbnails of all used source images are illustrated in Fig. 5.11. We utilized the combination scheme given in eqs. (4.10) and (3.52) for the DTCWT, the NSCT and the UWT-SF. As for the proposed fusion scheme with target extraction these rules got extended with the region-level

| Fusion metric | DTCWT | NSCT | UWT-SF | Proposed |
|---|---|---|---|---|
| $Q_{AB/F}$ | 0.5705 | 0.5757 | 0.6008 | **0.6021** |
| $Q_P$ | 0.7841 | 0.7899 | 0.7981 | **0.7995** |

**Table 5.1:** *Performance comparison of the achieved fusion metrics.*



(a) (b)

(c) (d)

**Figure 5.12:** *Fusion results of a sample image from the "UN Camp" sequence (frame 8). (a) UWT-SF fused. (b) UWT-SF with target extraction. (c) and (d) Zoomed versions of (a) and (b).*

fusion rules of Section 5.2. The objective evaluation of the obtained fusion results was accomplished by employing the $Q_{AB/F}$ and $Q_P$ fusion metrics as described in Section 3.2.

Table 5.1 lists the obtained fusion scores, averaged over all five tested IR-visible image pairs, for all tested fusion schemes using Piellas' match metric of eq. (5.4). It can be noticed that the UWT-SF as well as the proposed extension of the UWT-SF significantly outperform the fusion results obtained by state-of-the-art transforms such as the DTCWT and the NSCT for both fusion metrics. Note that this confirms once again the superiority of the proposed UWT-based fusion framework with spectral factorization of Chapter 4. Furthermore it can be seen that, by including target information into the fusion process, the fusion results of the UWT-SF can be further improved. This is most evident when looking at the fusion results of

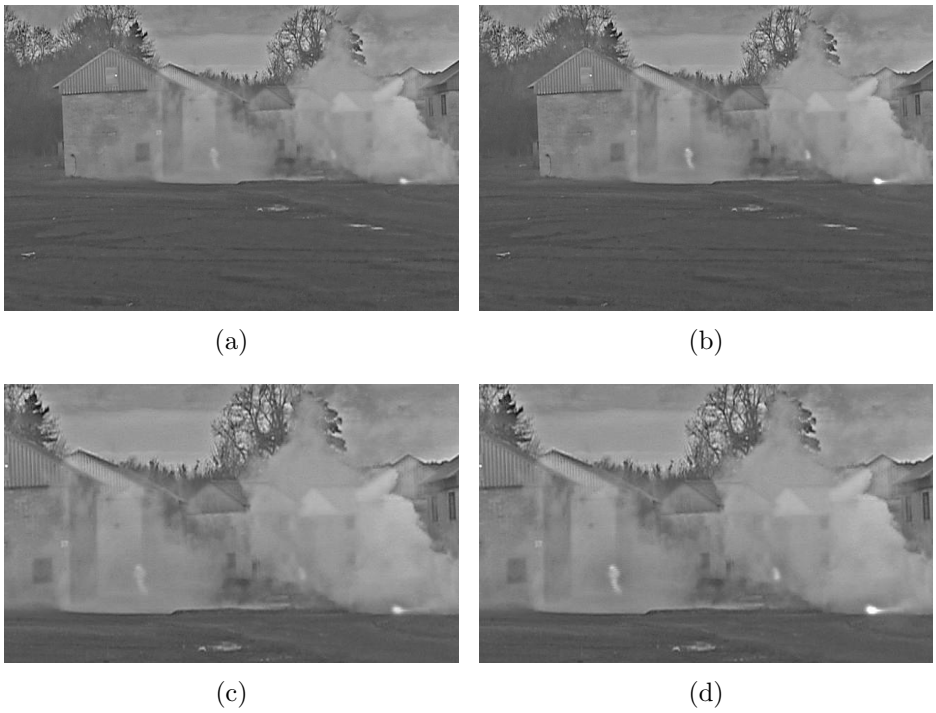


**Figure 5.13:** *Fusion results of a sample image from the "Octec" sequence (frame 21). (a) UWT-SF fused. (b) UWT-SF with target extraction and target enhancement. (c) and (d) Zoomed versions of (a) and (b).*

Fig. 5.12. It can be seen that the proposed extension produces fused images which show improved contrast and fewer ringing artifacts around target regions. This is particularly visible when observing the person depicted in the center of the zoomed images (Figs. 5.12(c) and (d)) which was correctly identified as an unambiguous target (not present in the visible image) by our target extraction algorithm.

Additionally, the proposed fusion method can be used to artificially "enhance" the extracted targets within the fused image. This is accomplished by multiplying all high-pass coefficients of the UWT belonging to a target region by a constant larger than one. The corresponding effect is shown in Fig. 5.13, where a multiplicative factor of 2 is used. Please note that this approach may lead to the introduction of additional artifacts in the fused image. However, due to the non-oscillating nature of the synthesis filter of eq. (4.9), these artifacts are not very disturbing.

Both tested match metrics were able to successfully distinguish between targets solely visible in the IR image and targets contained in both source images. However, for very small targets (e.g. second target from right in Fig. 5.8(a)) the S$\alpha$S model-based match metric exhibits unreasonable high differences between the target regions. We believe that this is due to the fact that their small number of pixels makes it difficult to extract meaningful model parameters from these regions, subsequently leading to unstable KLD values.

## 5.4 Conclusions

In this chapter an extension of the UWT-based pixel-level image fusion framework with spectral factorization of Chapter 4 is introduced that includes information about the presence of targets within the IR images into the fusion process. For this purpose a novel IR segmentation method was developed which is able to detect targets in low-contrast environments without introducing spurious results. Since the target extraction is performed on the decomposed images obtained after application of the first spectral factor, it can be embedded directly within the existing fusion framework. Additionally, a novel fusion scheme is proposed. It merges all extracted target regions assisted by a region-level fusion rule whereas the remaining image portions are fused using a pixel-level combination scheme. The usage of this hybrid approach turned the fusion process more robust against the introduction of segmentation-induced fusion errors, solving a classical problem of pure region-level fusion schemes.

We showed that our solution is able to improve the objective fusion results of the UWT-SF without target extraction, leading to a fused image with increased contrast and less reconstruction errors around target regions as verified by visual inspection. Furthermore, our proposed extension can be used to artificially enhance the visibility of the extracted targets within the fused image, supporting possible subsequent tasks such as target detection, localization and identification.

# Chapter 6

# A Novel Spatiotemporal IR/visible-light Video Registration Technique with Application to Image Fusion

Low production costs, increased resolution and high robustness of modern imaging sensors have made the use of multiple cameras common in many computer vision applications. Such multi-camera setups are particularly effective in environments where a single camera is incapable of capturing the entire information available within the monitored scene.

In this context, two different multiple camera setups can be identified. In the first one, cameras equipped with identical imaging sensors (e.g. visible-light sensors) are deliberately located at different viewpoints in order to increase the field-of-view of the overall imaging system. Applications for such multi-camera installations range from classical surveillance scenarios where one wishes to keep track of moving three dimensional (3D) objects as they move around a monitored area, to image stitching algorithms which are commonly used by the photogrammetry community to create high-resolution photo-mosaics [72].

The second setup starts from a different premise. Instead of observing a scene from different locations and combining the resulting views, the goal here is to produce a single image or video sequence containing information from various cameras positioned close to each other. Such imaging systems are of special interest since they allow one to exceed the physical bounds of a single sensor. In particular, they are utilized in spatial and temporal super-resolution frameworks (in order to improve the temporal and spatial resolution) as well as in image fusion applications for the purpose of increasing the overall depth of focus, the overall dynamic range

and the overall spectral response of the imaging system [124].

Independent of the underlying application, it is of vital importance that the utilized images are represented in a common reference coordinate frame. This can be achieved by jointly calibrating the employed cameras, that is, computing the optical properties (intrinsic parameters) as well as the relative positions of the individual cameras with respect to each other (extrinsic parameters). Based on these calibration parameters the images can subsequently be undistorted and rectified such that the pixel coordinates in one image sequence are in direct correspondence to pixel coordinates in the other image sequence. Please note that in the course of this work we will refer to this process as *image registration*.

Camera calibration methods can roughly be classified into traditional and self-calibration methods. Traditional calibration methods [125–133] usually require the cameras to simultaneously take several images of a calibration device. The actual calibration procedure then tries to localize a set of points within the calibration pattern of each view and computes the camera parameters based on these extracted calibration points. Typical choices of calibration points include the corners of a square pattern (checkerboard), the centers of circles, or the centers of a ring pattern [128]. Self-calibration methods on the other hand do not resort to the use of a calibration board. Instead they rely on the detection of a sufficient number of feature points within the source images (feature-point methods [124, 134–143]) or exploit common scene characteristics within the input images (direct methods [124, 144–147]) such as common illumination changes, appearance/disappearance of an object present in all image sequences, etc., to perform the calibration task.

In general, both traditional and self-calibration methods are well-suited for registering image sequences originating from cameras operating in the same spectral band. However, they tend to face problems for sequences obtained by sensors of different modalities (such as IR and visible-light sensors). For self-calibration methods this is mainly due to the possible lack of mutual feature-points or common scene characteristics within corresponding input images. The problems are less severe for traditional calibration methods. Nevertheless, the construction of a calibration board whose interest points appear likewise in the IR and visible-light spectrum and allow for the exact calibration of the employed cameras is not a trivial task.

As a consequence, only a few approaches to IR/visible-light stereo camera calibration can be found in the literature. Ukrainitz and Irani [145] introduce a self-calibration method for IR/visible-light video sequences based on maximizing local space-time correlations. In their work, affine transformations between corresponding image pairs are assumed. Other self-calibration methods for misaligned multi-sensor video sequences are presented in [136] and [124]. In [136], the authors assume a pair of IR/visible-light cameras to be jointly moved in space whereas in [124] a suffi-

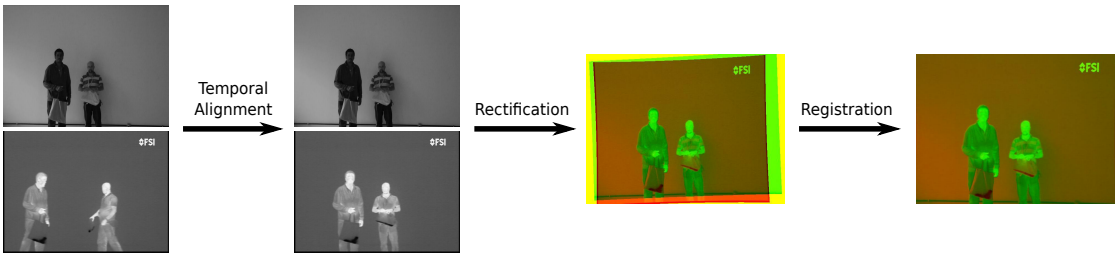



**Figure 6.1:** *Schematic diagram of the proposed IR/visible-light video registration framework. As for the superimposed pseudo-color images on the right, the visible-light and IR images occupy the green and red channels, respectively.*

cient number of mutual feature points needs to be tracked along the frames of an IR/visible-light video sequence pair. Traditional calibration methods for IR/visible-light stereo calibration include the ones in [130–133, 148]. However, these methods face problems extracting the precise calibration point positions from the images. Consequently, the mean reprojection error (MRE) - defined as the average error when mapping the calibration point positions from the world coordinate system to the image plane - of these methods is usually in the order of $10^{-1}$ [130, 131, 133] whereas state-of-the-art visible-light camera calibration approaches obtain a MRE of approximately $10^{-2}$ [128, 129].

In this work a novel IR/visible-light stereo camera calibration framework is introduced. The proposed approach uses a planar calibration board equipped with miniature light-bulbs to register a temporally and spatially misaligned IR/visible-light video sequence pair. Fig. 6.1 shows the schematic work flow of the proposed IR/visible-light video registration framework. In the course of this work we will show that the proposed system:

- is able to estimate the temporal offset between the IR and visible-light sequences in a very robust manner using solely the calibration point positions along both sequences, and
- leads to calibration results which exhibit significantly smaller MREs when compared to the state-of-the-art.

Finally, we will demonstrate the effectiveness of the proposed framework for image fusion, where co-registered images at sub-pixel accuracy are required. Please note that all registered IR/visible-light video sequence pairs are available for download at `http://www.smt.ufrj.br/~fusion/` and can be accessed freely by the research community. By doing so, we hope to alleviate the problem of most research in multimodal image fusion which suffers from an eminent lack of registered video sequences for evaluation purposes.

This chapter is structured as follows: Before introducing the proposed calibration point localization scheme in Section 6.2, the necessary background on the theory of

camera calibration is presented in Section 6.1. Based on the extracted calibration point positions, Section 6.3 introduces the overall temporal alignment approach whereas the proposed IR/visible-light camera calibration scheme is described in detail in Section 6.4. In Section 6.5 the experimental results obtained by applying the proposed framework to a number of temporally and spatially misaligned IR/visible-light video sequence pairs are presented. Finally, our conclusions are given in Section 6.6.

## 6.1 Background

In this section the main mathematical concepts involving camera calibration will be examined. For this purpose, we will first explain how 3D scene points can be accurately mapped onto a 2D image plane and derive the corresponding camera model. We will see that this projection can be represented by a $3 \times 4$ matrix together with a non-linear term which is used to correct the effects of lens distortion. Finally, based on the single camera model we will describe the epipolar geometry of two views and address the question how the knowledge of the position of an image point in one view constrains the position of the corresponding point in the other view.

In the course of this section the following notation will be used: Homogeneous 3D coordinates $\mathbf{X} = [X\ Y\ Z\ 1]^T$ will be represented by bold, capital letters whereas homogeneous 2D coordinates $\mathbf{x} = [x\ y\ 1]^T$ are given by bold, lowercase letters. Their inhomogeneous counterparts are denoted by $\tilde{\mathbf{X}} = [X\ Y\ Z]^T$ and $\tilde{\mathbf{x}} = [x\ y]^T$, respectively. As for stereo camera calibration, we will use the superscript $'$ to indicate entities associated with the second view.

### 6.1.1 Single Camera Calibration

Lets start our discussion with the basic pinhole camera model which is used in most computer vision applications to transform 3D world coordinates to 2D image coordinates. Let an image point in 2D be represented by the homogeneous vector $\mathbf{x}$ and its counterpart in the 3D world coordinate system by the homogeneous vector $\mathbf{X}$. As illustrated in Fig. 6.2, the general mapping given by the pinhole camera can be expressed by [1]

$$\mu\mathbf{x} = \mathbf{K}\,[\mathbf{R}\ \mathbf{t}]\,\mathbf{X}\ , \quad \text{with}\ \ \mathbf{K} = \begin{bmatrix} \alpha_x & s & x_0 \\ 0 & \alpha_y & y_0 \\ 0 & 0 & 1 \end{bmatrix}\ , \tag{6.1}$$

where $\mu$ is an arbitrary scale factor, $\mathbf{R}$ and $\mathbf{t}$ are the extrinsic camera parameters and $\mathbf{K}$ is called the intrinsic camera matrix [127] or camera calibration matrix [1].

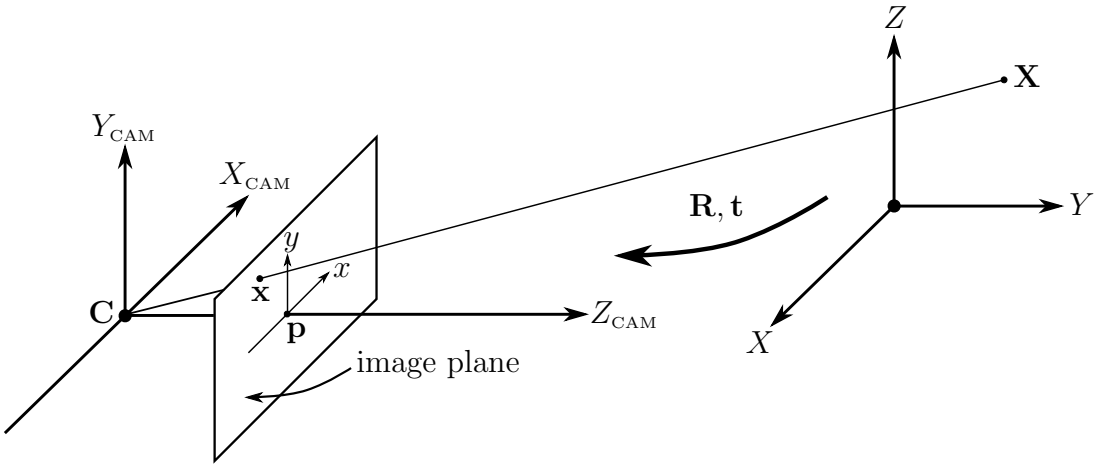



**Figure 6.2:** *Pinhole camera model. The mapping of a point* $\mathbf{X}$ *from the 3D world coordinate system to the point* $\mathbf{x}$ *in the 2D image coordinate system is given by* $\mathbf{x} = \mathbf{K}[\mathbf{R}\ \mathbf{t}]\mathbf{X}$ *where* $\mathbf{R}$ *and* $\mathbf{t}$ *define the Euclidean transformation between the world and camera coordinate system and* $\mathbf{K}$ *is the camera calibration matrix. The line from the camera center* $\mathbf{C}$ *perpendicular to the image plane is called the principal axis, and the point where the principal axis meets the image plane is called the principal point* $\mathbf{p}$.

The parameters of the $3 \times 3$ rotation matrix $\mathbf{R}$ and the $3 \times 1$ translation vector $\mathbf{t}$ represent the placement of the world coordinate system with respect to the camera coordinate system whereas $\mathbf{K}$ contains the internal camera parameters in terms of pixel dimensions. These are the focal length $(\alpha_x, \alpha_y)$ and the principal point $(x_0, y_0)$ of the camera in the $x$ and $y$ direction, respectively, as well as the parameter $s$ which describes the skewness of the two image axes.

Two particularly important classes of camera matrices can be derived from the camera model of eq. (6.1): finite cameras, and cameras with their center at infinity (such as the affine camera which represents parallel projection) [1]. In this work we will mainly focus on finite cameras corresponding to the set of homogeneous $3 \times 4$ matrices $\mathbf{P} = \mathbf{K}[\mathbf{R}\ \mathbf{t}]$ for which the left hand $3 \times 3$ submatrix $\mathbf{KR}$ is non-singular.

As it is rather difficult to make a good 3D calibration device, one often uses multiple views of a planar calibration pattern for calibration purposes. When using such a calibration device we can assume without loss of generality that the calibration pattern is located on the plane $Z = 0$ in the world coordinate system. Thus, we can rewrite eq. (6.1) such that

$$\mu \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \mathbf{K}\left[\mathbf{r_1}\ \mathbf{r_2}\ \mathbf{r_3}\ \mathbf{t}\right] \begin{bmatrix} X \\ Y \\ 0 \\ 1 \end{bmatrix} = \mathbf{K}\left[\mathbf{r_1}\ \mathbf{r_2}\ \mathbf{t}\right] \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} = \mathbf{H}\bar{\mathbf{X}}\ , \tag{6.2}$$

where $\mathbf{R}$ is given by $[\mathbf{r_1}\ \mathbf{r_2}\ \mathbf{r_3}]$, $\mathbf{H} = \mathbf{K}\left[\mathbf{r_1}\ \mathbf{r_2}\ \mathbf{t}\right]$ is called a homography matrix and $\bar{\mathbf{X}} = [X\ Y\ 1]^T$.

As shown in [127], by knowing the homographies $\mathbf{H}$ between the calibration pattern and its image for two or more views, a first estimate of the intrinsic and extrinsic parameters can be obtained by applying a closed-form algorithm such as the Direct Linear Transformation (DLT) algorithm. However, this first estimate of the calibration parameters is not optimal since a) it is obtained by applying the singular value decomposition which minimizes an algebraic distance measure that is not physically meaningful, and b) does not consider radial and tangential distortion arising from the optical lens employed in the camera. Lens distortion can be incorporated using the following expression [125, 126]

$$\mathcal{F}(\tilde{\mathbf{x}}_c, \mathcal{K}, \mathcal{P}) = \begin{bmatrix} x_c \left(k_1 r^2 + k_2 r^4 + \ldots\right) + \left(2p_1 x_c y_c + p_2(r^2 + 2x_c^2)\right) \\ y_c \left(k_1 r^2 + k_2 r^4 + \ldots\right) + \left(p_1(r^2 + 2y_c^2) + 2p_2 x_c y_c\right) \end{bmatrix}, \tag{6.3}$$

where $\tilde{\mathbf{x}}_c = [x_c \; y_c]^T$ are the (non-observable) distortion-free, normalized points in the camera coordinate system before applying the camera calibration matrix $\mathbf{K}$, $\mathcal{K} = \{k_1, k_2, \ldots\}$ and $\mathcal{P} = \{p_1, p_2\}$ are the coefficients of the radial and tangential distortion, respectively, and $r = \sqrt{x_c^2 + y_c^2}$. The (observable) distorted, normalized points $\tilde{\mathbf{x}}_d$ are then approximated by

$$\tilde{\mathbf{x}}_d = \tilde{\mathbf{x}}_c + \mathcal{F}(\tilde{\mathbf{x}}_c, \mathcal{K}, \mathcal{P}) \tag{6.4}$$

and the final image points are given by $\mathbf{x} = \mathbf{K}\mathbf{x}_d$. Note that in this work a $2^{\text{nd}}$ order radial distortion model with tangential distortion is used such that $\mathcal{K} = \{k_1\}$ and $\mathcal{P} = \{p_1, p_2\}$.

With all this in mind, a final global optimization step is incorporated which estimates the complete set of parameters using the previously obtained calibration parameters as an initial guess. This optimization is done iteratively by minimizing the following functional [127]

$$\sum_i \sum_j \left\| \mathbf{x}_{ij} - \breve{\mathbf{x}}\left(\mathbf{K}, \mathcal{K}, \mathcal{P}, \mathbf{R}_i, \mathbf{t}_i, \bar{\mathbf{X}}_j\right) \right\|^2, \tag{6.5}$$

where $\mathbf{x}_{ij}$ is the sub-pixel position of the $j^{\text{th}}$ calibration point in the $i^{\text{th}}$ calibration image, and $\breve{\mathbf{x}}\left(\mathbf{K}, \mathcal{K}, \mathcal{P}, \mathbf{R}_i, \mathbf{t}_i, \bar{\mathbf{X}}_j\right)$ is the projection of the corresponding calibration point $\bar{\mathbf{X}}_j$ from the 3D world coordinate system.

Given the calibration point positions in the real world and camera coordinate system, various off-the-shelf solutions for camera calibration exist. Among them, the *OpenCV Camera Calibration Toolbox* [149] as well as the *Camera Calibration Toolbox for Matlab* [150] are predominately used.

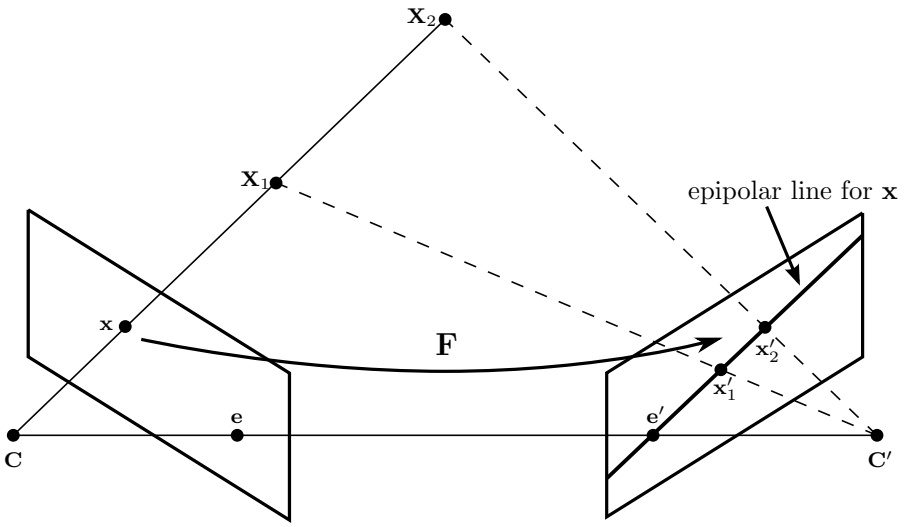



**Figure 6.3:** *An image point* $\mathbf{x}$ *in the left view back-projects to a ray in the 3D world coordinate system. This ray is imaged as a line in the right view. All points located on the ray are imaged at* $\mathbf{x}$ *in the left view whereas they generate distinct image points in the right view. The epipoles* $\mathbf{e}$ *and* $\mathbf{e}'$ *are the points of intersection of the line joining the camera centers* $\mathbf{C}$ *and* $\mathbf{C}'$ *(camera baseline) with the image planes. Corresponding points* $\mathbf{x} \leftrightarrow \mathbf{x}'$ *satisfy the constraint* $\mathbf{x}'^T\mathbf{F}\mathbf{x} = 0$*, where* $\mathbf{F}$ *is called the fundamental matrix of the camera pair.*

### 6.1.2 Stereo Camera Calibration

In this subsection we formally define the epipolar geometry between a pair of images. As before, we will start with the basic pinhole camera model which does not assume lens distortion. Suppose a 3D scene point $\mathbf{X}$ is imaged at the point $\mathbf{x}$ in the first view and at $\mathbf{x}'$ in the second view. Then corresponding points $\mathbf{x} \leftrightarrow \mathbf{x}'$ satisfy the following epipolar constraint [1]

$$\mathbf{x}'^{\mathrm{T}}\mathbf{F}\mathbf{x} = 0, \tag{6.6}$$

where $\mathbf{F}$ is called the fundamental matrix of the camera pair. An important property of the fundamental matrix is that it is of rank 2. As a consequence $\mathbf{F}$ does not provide point-to-point correspondences. Instead it specifies a map $\mathbf{x} \mapsto \mathbf{l}'$ from a point in one image to its corresponding epipolar line in the other image, as illustrated in Fig. 6.3.

If we assume that both cameras, represented by the matrices $\mathbf{P}$ and $\mathbf{P}'$, have been calibrated according to the pinhole camera model such that

$$\mathbf{P} = \mathbf{K}\,[\mathbf{I} \mid \mathbf{0}] \qquad \mathbf{P}' = \mathbf{K}'\,[\mathbf{R} \mid \mathbf{t}]\,, \tag{6.7}$$

where, without loss of generality, we choose the world origin to coincide with the

first camera $\mathbf{P}$, then the fundamental matrix can be expressed by [1]

$$\mathbf{F} = [\mathbf{K}'\mathbf{t}]_{\times} \mathbf{K}'\mathbf{R}\mathbf{K}^{-1}, \tag{6.8}$$

where we use the notation that the 3-vector $[\mathbf{K}'\mathbf{t}]_{\times}$ defines a $3 \times 3$ skew-symmetric matrix such that the vector product $\mathbf{a} \times \mathbf{b} = [\mathbf{a}]_{\times}\mathbf{b}$, and $\mathbf{R}$ and $\mathbf{t}$ describe the relative rotation and displacement of the two cameras, respectively.

Due to the linearity of eq. (6.8), the fundamental matrix provides a simple and computationally friendly solution to compute point-to-line correspondences within a stereo camera setup. However, for real cameras employing optical lenses such a linear mapping is no longer valid. To this end, the mapping of image points from the first view to the second view in the presence of lens distortion can be summarized as follows: First, apply the inverse camera calibration matrix to the 2D image points in the first view $\mathbf{x}_d = \mathbf{K}^{-1}\mathbf{x}$. Next, in order to obtain the distortion-free, normalized points $\mathbf{x}_c$, the inverse distortion model of eq. (6.3) needs to be employed to $\mathbf{x}_d$. However, this is not straightforward since no analytic solution for the inverse exists. One way to bypass this problem is to approximate the inverse distortion model recursively [125, 126]

$$\begin{aligned} \tilde{\mathbf{x}}_c &\approx \tilde{\mathbf{x}}_d - \mathcal{F}(\tilde{\mathbf{x}}_d, \mathcal{K}, \mathcal{P}) \approx \tilde{\mathbf{x}}_d - \mathcal{F}(\tilde{\mathbf{x}}_d - \mathcal{F}(\tilde{\mathbf{x}}_d, \mathcal{K}, \mathcal{P}), \mathcal{K}, \mathcal{P}) \\ &\approx \tilde{\mathbf{x}}_d - \mathcal{F}(\tilde{\mathbf{x}}_d - \mathcal{F}(\tilde{\mathbf{x}}_d - \mathcal{F}(\tilde{\mathbf{x}}_d, \mathcal{K}, \mathcal{P}), \mathcal{K}, \mathcal{P}), \mathcal{K}, \mathcal{P}) \approx \dots \ . \end{aligned} \tag{6.9}$$

By doing so, the error introduced when substituting $\mathbf{x}_d$ with $\mathbf{x}_c$ on the right-hand side gets smaller for each iteration. As shown in [125, 126] three to four iterations are sufficient to compensate for strong lens distortions. As a next step, the undistorted points $\mathbf{x}_c$ are mapped from the first camera coordinate system through the plane at infinity [1] to the camera coordinate system of the second camera [1] ($\mathbf{x}'_c = \mathbf{R}\mathbf{x}_c$) and lens distortion is added using the forward lens distortion model of eq. (6.3) such that $\tilde{\mathbf{x}}'_d = \tilde{\mathbf{x}}'_c + \mathcal{F}(\tilde{\mathbf{x}}'_c, \mathcal{K}', \mathcal{P}')$. Finally, by applying the camera calibration matrix ($\mathbf{x}' = \mathbf{K}'\mathbf{x}'_d$), a potential match of $\mathbf{x}$ in the second view is found. Please note that, as a consequence of lens distortion, the previously established point-to-line correspondences does not hold anymore. Instead, if points $\mathbf{x}$ and $\mathbf{x}'$ correspond, then $\mathbf{x}'$ lies on a curved epipolar line controlled by the polynomial distortion function of eq. (6.3).

Apart from the chosen camera model, the overall accuracy of camera calibration depends to a great extent on the ability to localize the set of calibration points within the provided calibration footage. Thus, in the next section we will introduce a novel calibration point localization scheme which is able to find the set the calibration points with very high accuracy.

## 6.2 Calibration Point Detection

Due to the different spectral sensitivity of IR and visible-light cameras, the construction of a calibration board whose interest points appear both in the visible-light and IR spectra is not a trivial task. For example, existing camera calibration approaches based on black/white calibration patterns cannot be employed straightforwardly since such calibration devices do not appear in the IR image in most cases.

As a consequence, only a few calibration devices have been proposed in the literature for IR/visible-light camera calibration. Prakash et al. [148] advocate a heated chessboard as an appropriate calibration device. The authors argue that due to the different IR emissivity of black and white regions, it is possible to extract the corner points of the chessboard pattern in the visible-light and IR modality, respectively, and use these points for calibration purposes. However, as reported in [133], such a calibration board fails to exhibit crisp corners in the IR spectrum, consequently preventing the precise localization of these corner points in the IR image. Thus they suggest a different calibration pattern which consists of a grid of regularly sized squares, cut out of a material that is opaque in the IR modality. The authors demonstrate that when held in front of a backdrop with a different temperature than the ambiance, such a pattern can be identified in the IR domain and allows for a more reliable extraction of the corner points. Another interesting strategy is chosen in [131], where a planar black/white checkerboard pattern is augmented by a set of resistors mounted in the centroid of each square. In this approach, the corners of the black/white squares are utilized for the calibration of a visible-light camera, whereas the energized resistors are used for IR camera calibration.

The calibration board chosen in this work uses miniature light bulbs, equidistantly mounted on a planar calibration board [130, 132]. This configuration is of special interest since, when energized, heat and light are simultaneously emitted by the light bulbs causing the calibration pattern to appear in both the visible-light and IR modalities. This is demonstrated in Fig. 6.4, where the employed calibration board consisting of 81 light bulbs, arranged in a $9 \times 9$ matrix is shown in the visible-light and IR spectrum, respectively. Please note that the depicted images were taken from an IR/visible-light video sequence after successful temporal alignment.

The main advantages of the chosen calibration board include its versatility (e.g. the calibration board can be used for daytime and nighttime recordings), its fast operational readiness ("plug & play") and its easy portability. Moreover, since the same physical entities (light bulbs) are used as calibration points in the IR/visible-light images, eventual imperfections of the calibration board (e.g. loose contact of one of the light bulbs) can be compensated more easily.


(a) (b)

**Figure 6.4:** *Employed calibration board consisting of 81 light bulbs, arranged in a* $9 \times 9$ *matrix, in the (a) visible-light and (b) IR spectrum. The depicted images were taken from an IR/visible-light image sequence after temporal alignment.*

Nevertheless, when observing Fig. 6.4 some challenges associated with the chosen calibration board can be identified. For one thing, due to the use of cheap, off-the-shelf light bulbs, the emitted radiation patterns tend to differ from light bulb to light bulb - a problem which is further aggravated when tilting the calibration board. In extreme cases, this may even lead to the fading of some light bulbs. For another thing, the visibility of the light bulbs in the visible-light image depends to a high extend on the surrounding lightning conditions. For example, for outdoor sequences recorded at bright day light, the calibration points are less noticeable than for indoor scenes where the lightning conditions can be controlled.

In order to cope with these challenges, a series of steps are proposed to robustly extract the sub-pixel positions of the miniature light bulbs along all video frames exhibiting the calibration board of Fig. 6.4. An overview of the proposed algorithm is given in Algorithm 1. In the remainder of this section the individual steps will now be detailed.

### 6.2.1 Calibration Board Detection

For visible-light calibration footage, the localization of the light bulb regions often poses difficulties. Especially in day light scenes recorded at bright lightning conditions the used miniature light bulbs tend to be poorly visible due to their limited illumination capacities. As a consequence, the localization of the light bulb regions without pre-processing of the calibration images may result in a high number of false positives and in the worst case even to detection failures. A possible way to alleviate this problem is to restrict the search region to the area constrained by the edges of the calibration board. As can be observed in Fig. 6.4(a), this approach seems promising since the borders of the calibration board are clearly noticeable in the visible-light calibration images.

In order to locate the pattern, we first segment the input images into a set of

**Objective**

Find the sub-pixel positions of the calibration points in each image exhibiting the calibration device of Fig. 6.4.

**Algorithm**

1) **Detect the calibration board** (*only visible-light images*):
   (a) Apply marker-controlled Watershed Transformation to the calibration images.
   (b) Use available ground truth about the calibration board to remove wrongly-extracted regions:
      - Overall size of the region is less than a threshold.
      - Region is not square-shaped.
      - In case of more than one region satisfying the above conditions, choose the one which contains the most sub-regions within its borders.
   (c) Utilize the Hough transform to obtain the final borders of the calibration board.

2) **Find the positions of the calibration points**:
   (a) Define an initial threshold value $\lambda$.
   (b) Perform gray-scale thresholding to separate the light bulb regions from the background.
   (c) Fit an ellipse to each extracted region and compute its centroid.
   (d) Based on the eccentricity of the fitted ellipses, remove regions which are not circular.
   (e) If the number of regions is smaller than the number of light bulbs:
      - Increase the value of $\lambda$ and go to step (b).

      If the number of regions is higher than the number of light bulbs:
      - For each region $\mathcal{R}_i$, calculate the average distance $d_i$ to its two closest neighbors.
      - Compute the median $\tilde{d}$ over the whole set of distance measures $d_i$, where $i = 1, \ldots, N$
      - Remove those regions $\mathcal{R}_i$ whose corresponding distance measure $d_i$ differs most from the median distance $\tilde{d}$. The number of regions eliminated this way is chosen such that the combinatorial complexity for the step below is reduced to an applicable degree.
      - Randomly choose a set of candidate regions (corresponding to the number of light bulbs) and compute the MRE of eq. (6.10). Repeat this for all possible combinations and choose the set for which the MRE is a minimum.
   (f) Use the DLT algorithm in conjunction with the computed centroids to determine a first estimate of the homography matrix $\mathbf{H}$ and compute the resulting MRE.
   (g) If the MRE is greater than some constant $\varepsilon$, increase the value of $\lambda$ and try again from step (b).
   (h) Refine the computed homography by minimizing eq. (6.11).
   (i) Compute the final calibration point positions by applying the refined homography to the calibration point positions in the world coordinate system.

**Algorithm 1:** Proposed calibration point localization scheme.

(a) (b)

**Figure 6.5:** *Results of the calibration board detection for the visible-light calibration image of Fig. 6.4(a). (a) Segmentation result of the marker-controlled watershed transformation. (b) Detected calibration board after application of the Hough transform.*

candidate regions encompassing the sought calibration pattern. Since the calibration board forms a bounded region within the calibration images, a convenient way to do so is to use the watershed transformation which, when applied to the gradient image, partitions the original gray-scale image into regions of homogeneous gray-scale values. However, as pointed out in Chapter 5, the direct application of the watershed transformation leads to a considerable over-segmentation of the input image due to noise or eventual local gray-scale oscillations of the gradient image. One way to improve the result is the use of the watershed transformation together with a marker image, limiting the segmentation process to some areas of interest. In this work, the marker image is obtained by applying the Canny edge detector [151] to the calibration images and by following the processing chain given in Section 5.1.1.

Before performing the marker-controlled watershed transformation it is advantageous to simplify the calibration images. As outlined in Section 5.1.2, this is accomplished by computing the morphological gradient of the calibration images and by quantizing the result to 100 gray levels. Once the simplified image is obtained, it is combined with the marker image and the watershed transformation is computed. Fig. 6.5(a) illustrates the result of this process for the visible-light calibration image of Fig. 6.4(a).

However, as can be seen in Fig. 6.5(a), the segmented image still contains a significant amount of over-segmentation. These wrongly-extracted regions are removed using the available ground-truth about the calibration board. More specifically, we discard regions a) whose overall size is less than a certain threshold or b) which are not square-shaped, assessed by calculating the ratio between the regions' perimeter and its area. If there should be more than one candidate region satisfying both conditions, the number of contained sub-regions is utilized as a tie-breaker.

Finally, in order to turn the extracted calibration board region into a true rect-

angle with well-defined vertices, we apply the Hough transform [152] and extract the four most prominent features in Hough space. Fig. 6.5(b) shows the detected calibration board within the visible-light calibration image of Fig. 6.4(a) after application of the Hough transform.

It is worth repeating that henceforth all operations applied to visible-light calibration images with known calibration board location will be confined to the calibration board region.

### 6.2.2 Calibration point localization

In order to compute the exact sub-pixel positions of the miniature light bulbs along all IR/visible-light video frames exhibiting the calibration board of Fig. 6.4, we first have to separate the light bulb regions from the background. Ideally, this would be accomplished by applying a static threshold to the calibration images, labeling all pixels above the threshold as belonging to a potential light bulb region. However, due to the varying appearance of the light bulbs, no global threshold is capable of reliably producing a binary image that contains all light bulbs whilst suppressing the number of falsely extracted background regions.

Thus, the approach taken in this work does not rely on a single global threshold but tries to extract the exact light bulb positions by iteratively determining the optimal threshold for each calibration image. For this purpose, we first choose an initial threshold (either manually or by means of some adaptive thresholding scheme as the one in [129]) which is subsequently used to binarize the calibration image. After the thresholding operation, the extracted light bulb regions are expected to exhibit ellipse-like patterns in the binarized image. Based on this assumption, we post-process the binary image by removing all regions which appear with arbitrary shape and do not resemble the expected ellipsoidal radiation pattern. This is accomplished by fitting an ellipse to the boundary pixels of each region and discarding those for which the committed error (defined as the sum of squares of the distances between the boundary pixels of the region and the fitted ellipse) is above some threshold. Furthermore, we also remove those regions corresponding to ellipses with high eccentricity (measure of how much the ellipse deviates from being circular) since it is assumed that the ellipses corresponding to light bulb regions closely resemble a circle. Note that in our implementation the ellipse fitting is performed by employing the algorithm of [153].

A first estimate of the calibration point positions is obtained by substituting the original light bulb regions with the area of the computed ellipses and by calculating its centroids within the original calibration images. If the number of computed calibration points is below the overall number of light bulbs we repeat the above

procedure using the next higher threshold. If on the other hand the number of extracted calibration points is higher than the number of light bulbs, a potential solution is to randomly choose a subset of calibration points from the complete set and to compute the corresponding homography using the DLT algorithm. If the correct set was chosen, mapping the calibration points from the 3D world coordinate system to the calibration image will be quite precise and will consequently result in a small MRE, defined as

$$\mathrm{MRE} = \frac{1}{N} \sum_{i} \left\| \mathbf{x}_i - \mathbf{H}\bar{\mathbf{X}}_i \right\| . \tag{6.10}$$

Here, $N$ is the total number of light bulbs, $\mathbf{x}_i$ is the estimated position of the $i^{\text{th}}$ calibration point within the calibration image and $\bar{\mathbf{X}}_i$ represents the position of the corresponding calibration point in the world coordinate system. On the contrary, if the MRE is high we have strong evidence that the chosen subset does not correspond to the true light bulb positions and another subset needs to be chosen.

Even though this procedure was found to be very robust, it is computationally expensive when the number of extracted regions is much larger than the actual number of light bulb regions. In fact, in a scenario with $k$ light bulbs and $n$ extracted regions with $n > k$ the combinatorial complexity of this approach can be expressed by the binomial coefficient $\binom{n}{k}$ resulting in $\frac{n!}{k!(n-k)!}$ different combinations. It is easy to verify that the number of possible combinations grows exponentially with the number of extracted regions. For instance, for the case of 81 light bulbs and 82, 83 and 84 extracted regions, respectively, the overall number of combinations is 82, 3403 and 95284.

Thus, in situations where the ratio of extracted calibration points to light bulbs renders the above mentioned method impracticable, a preliminary step for outlier removal is needed. This is done by exploiting the available information about the light bulb distribution on the calibration board. In more detail, assuming that the distances between pairs of adjacent light bulbs are approximately constant within the calibration images, we iteratively eliminate the calibration points whose mean distance to its closest neighbors differ most from the median distance, calculated over the whole set of extracted regions. This procedure is repeated until the combinatorial complexity for the aforementioned method is reduced to an acceptable degree such that it can be used to remove all remaining falsely extracted calibration points without causing a high computational overhead.

If the number of extracted calibration points matches the number of light bulbs, and the corresponding MRE is below a pre-defined threshold, then the final calibration point positions can be computed. Since our goal is to obtain calibration points

(a) (b)

**Figure 6.6:** *Results of the calibration point detection for the (a) visible-light and (b) IR calibration images of Fig. 6.4 (zoomed version).*

for which the MRE is a minimum we refine the corresponding homography $\mathbf{H}$ by minimizing the functional

$$\min_{\mathbf{H}} \sum_{i} \left\| \mathbf{x}_i - \mathbf{H}\bar{\mathbf{X}}_i \right\| . \tag{6.11}$$

The final calibration point positions are computed by applying the refined homography to the calibration point positions in the world coordinate system. Fig. 6.6 shows the resulting calibration point positions for the visible-light and IR calibration image of Fig. 6.4(a) and Fig. 6.4(b), respectively.

It will be shown in the next section that by means of the extracted calibration point positions, the time-shift between two unsynchronized IR and visible-light sequences can be determined successfully.

## 6.3 Temporal Alignment

Let $\mathbf{S}_V$ and $\mathbf{S}_I$ be two video sequences $N_V$ and $N_I$ frames long, recorded at the same frame rate by a visible-light and an IR camera, respectively, exhibiting different poses of the calibration board of Fig. 6.4. Finding the temporal offset $\Delta\hat{t}$ between the two video sequences $\mathbf{S}_V$ and $\mathbf{S}_I$ is equivalent to maximizing a similarity measure $s(\cdot)$ over a set of potential temporal offset candidates $\Delta t$ such that

$$\Delta\hat{t} = \arg\max_{\Delta t} \; s\big(\mathbf{S}_V, \mathbf{S}_I, \Delta t\big). \tag{6.12}$$

Please note that, in what follows, we start from the premise that the two video cameras are mounted side by side on an horizontal rail. Therefore, we assume that their positions only differ horizontally and are identical otherwise.

The proposed temporal alignment approach starts off by performing transla-

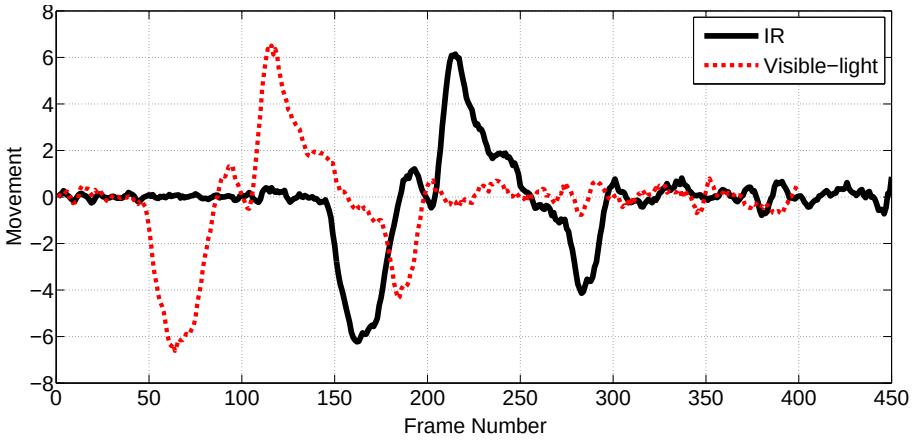
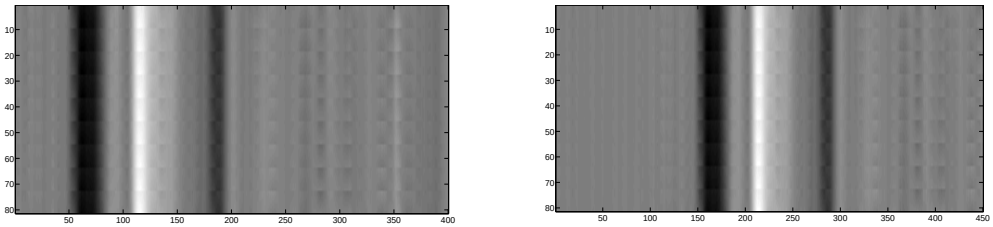



**Figure 6.7:** *Example of the vertical component of the speed of a single calibration point along a visible-light (dashed line) and an IR (solid line) video sequence.*



(a) (b)

**Figure 6.8:** *Global movement of all* 81 *calibration points along a (a) visible-light and (b) IR video sequence. Each line represents the vertical movement of a single calibration point. Bright pixel values indicate an upward movement whereas dark pixel values represent a downward movement of the calibration board.*

tional movements of the calibration board in the downward and upward direction, respectively. This is followed by the extraction of the calibration point positions in each frame of the IR and visible-light video sequence as elaborated in Section 6.2. Based on the extracted calibration point positions, we determine the vertical component of the speed of each calibration point along the video sequences. This is accomplished by subtracting the $y$-coordinates of the calibration point positions between two successive video frames. Fig. 6.7 shows an example of the vertical speed of a single feature point along an IR/visible-light video sequence pair. From the depicted curves the downward and upward swing of the calibration board, given by the negative and positive portions of the curves, respectively, can be seen.

Another way to look at the problem at hand is presented in Fig. 6.8. Here, the global movement of all calibration points is represented as an image with each line representing the overall vertical movement of a single calibration point. In both images, brighter pixel values indicate the displacement of the calibration board in the upward direction whereas darker pixel values suggest a downward movement of the calibration pattern. Based on Fig. 6.8, the temporal offset between the two

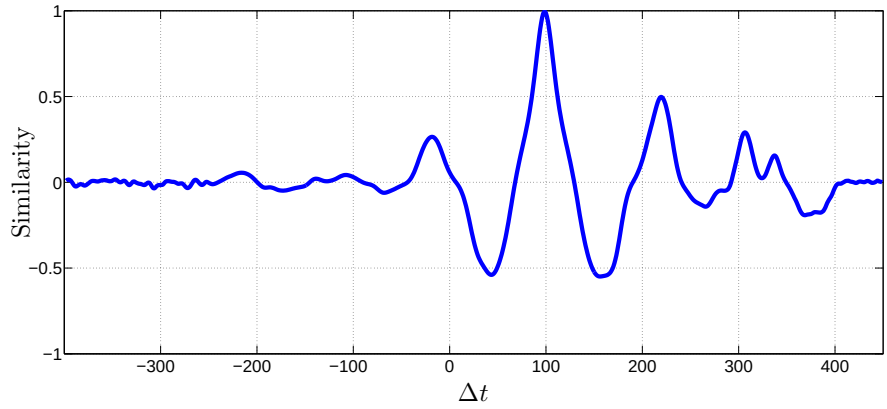



**Figure 6.9:** *Result of the temporal alignment for the two IR and visible-light video sequences corresponding to Fig. 6.8. The highest similarity (according to eq. (6.13)) between the two video sequences is obtained for a temporal offset* $\Delta t$ *of* 99 *frames.*

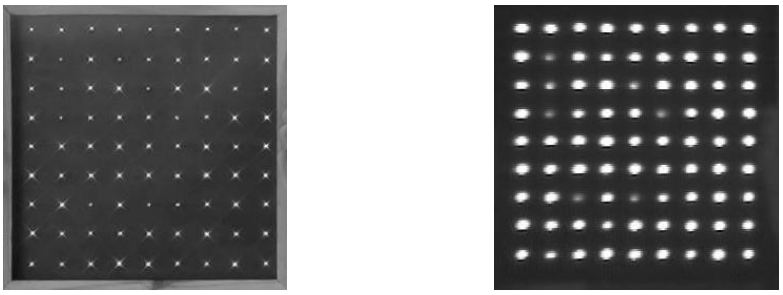
video sequences can be described straightforwardly. It simply corresponds to the horizontal displacement between the two images for which their horizontal cross-correlation is maximized.

Let us put this observation now in a mathematical context. Given a temporal offset candidate $\Delta t$, the similarity between the visible-light sequence $\mathbf{S}_V$ and the IR sequence $\mathbf{S}_I$ is given by

$$s(\mathbf{S}_V, \mathbf{S}_I, \Delta t) = \frac{\sum_{m=1}^{M} \sum_{n \in \mathcal{N}} \mathbf{M}_V(m, n - \Delta t) \mathbf{M}_I(m, n)}{\sqrt{\sum_{m=1}^{M} \sum_{n \in \mathcal{N}} \big(\mathbf{M}_V(m, n - \Delta t)\big)^2 \sum_{k=1}^{K} \sum_{l \in \mathcal{N}} \big(\mathbf{M}_I(k, l)\big)^2}}, \tag{6.13}$$

where the matrices $\mathbf{M}_V(m, n)$ and $\mathbf{M}_I(m, n)$ express the movement of the $m^{\text{th}}$ calibration point between two consecutive visible-light and IR frames at time instant $n$, respectively, and $\mathcal{N} = \{n \mid 1 \leq (n - \Delta t) \leq N_V \wedge 1 \leq n \leq N_I\}$. Please note that the similarity measure of eq. (6.13) is bounded to the interval $[-1, 1]$. The two video sequences are considered identical if the similarity measure is 1 and complementary to each other if the result is $-1$. A result of 0 implies that no similarities between the two sequences could be found. Finally, as expressed in eq. (6.12), the true temporal offset $\Delta\hat{t}$ between the IR and visible-light video sequence is the one for which eq. (6.13) is maximized.

Please note that due to the discrete sampling over time, the offset mostly falls in between two frames. As such it would be necessary to temporally resample each single frame within one video sequence, creating a considerable computational overhead. In the presented work this is avoided by approximating the time-shift by its closest integer value, resulting in a trade-off between synchronization precision



(a) (b)

**Figure 6.10:** *Undistorted views of the calibration boards of Fig. 6.4 in the fronto-parallel plane. (a) Visible-light image. (b) IR image.*

and computational complexity.

Fig. 6.9 shows the result of the temporal alignment for the two IR/visible-light video sequences corresponding to Fig. 6.8. It can be observed that the highest similarity (according to eq. (6.13)) is obtained for a temporal offset of 99 frames. This result corresponds well with Fig. 6.8 which, when evaluated subjectively, suggests a time-shift of approximately 100 frames between the two sequences.

## 6.4 Camera Calibration

Once the IR/visible-light video sequence pair is synchronized, the individual and joint camera parameters of the IR/visible-light camera pair can be estimated. This is accomplished by choosing $N$ temporally aligned calibration images and by following the calibration procedure outlined in Section 6.1.1. Please note that in the current implementation the calibration images where chosen manually such that a high variety of different poses of the calibration board is incorporated in the calibration process. However, in a future version this process may be automated by extracting the pose information directly from the homography matrices [154, 155].

A major limitation of the chosen calibration approach is that the calibration point localization as described in Section 6.2 is performed using non-fronto-parallel calibration images which suffer from nonlinear distortions due to the camera optics. In order to improve calibration results, it is therefore beneficial to first map the calibration images onto an undistorted fronto-parallel view (see Fig. 6.10) and determine the exact calibration point positions within these canonical images. However, in order to do so, full knowledge of the calibration parameters would be necessary - information that is usually not available at this point. One possible solution to this problem is presented in [128] where the authors advocate an iterative refinement approach, using alternating mappings of the calibration images onto a canonical fronto-parallel view and back.

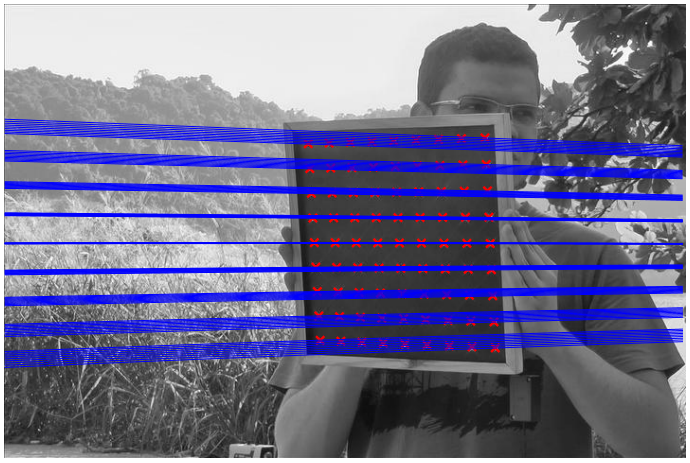


**Figure 6.11:** *Result of stereo calibration when mapping the IR calibration points of Fig. 6.6(b) to the visible-light calibration image of Fig. 6.6(a). Note that due to lens distortion this mapping is no longer linear, resulting in curved epipolar lines.*

In this work we follow a similar approach. After calculating a first preliminary version of the calibration parameters we remove the radial and tangential distortion from the calibration images and map them onto a canonical fronto-parallel plane in the world coordinate system. Within this fronto-parallel view we then localize the calibration points using the processing chain of Section 6.2. Finally, these new calibration points are remapped onto the original image plane and the camera parameters are recomputed using the updated calibration point positions. This process is repeated until convergence, where in each new loop the mapping onto the fronto-parallel plane is performed using the camera parameters from the previous iteration. Fig. 6.10 shows the undistorted equivalents of Fig. 6.4 in the fronto-parallel view. As will be shown in Section 6.5, the calibration parameters obtained by means of this iterative calibration point refinement result in a reprojection accuracy exceeding the one of traditional IR/visible-light camera calibration approaches.

After completing the individual calibration procedures for the IR and visible-light camera we jointly calibrate them as described in Section 6.1.2. By doing so, we gain knowledge of the relative displacement of the two cameras, consequently enabling us to map points from one view to the other one. As previously pointed out, due to lens distortion this mapping is not linear in the sense that a point in one view does not induce a line in the other view. Instead a curved line is generated on which the corresponding point in the second view resides. This is demonstrated in Fig. 6.11 where the epipolar curves resulting from mapping the IR calibration points of Fig. 6.6(b) to the visible-light calibration image of Fig. 6.6(a) are highlighted. It can be observed that the distances between the epipolar curves and the corresponding calibration points are very small, suggesting a high accuracy of the stereo calibration results.

**Figure 6.12:** *Result of image rectification for a sample IR/visible-light image pair. For visualization purposes, the two images where overlaid on top of each other and occupy the red (visible-light) and green (IR) channel within the depicted RGB pseudo-color image.*

Next, based on the obtained epipolar geometry we rectify the IR/visible-light image pairs [1, 156], resulting in image correspondences where the epipolar curves are linearized and run parallel to the $x$-axis. By doing so, disparities between the IR/visible-light image pairs will occur in the $x$-direction only. As a consequence, rectification may be used to recover 3D structure information by providing the depth discrepancies between the rectified image pairs. In this work rectification is achieved by undistorting both image sequences using eq. (6.9) and applying two rectifying homographies $\mathbf{H}_R$ and $\mathbf{H}'_R$ to the undistorted IR and visible-light images, respectively, such that, after rectification, point correspondences are given by [156]

$$\left(\mathbf{x}'^T \mathbf{H}_R'^T\right) \mathbf{F} \left(\mathbf{H}_R \mathbf{x}\right) = 0 \qquad \text{where} \qquad \mathbf{F} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix} \tag{6.14}$$

and $\mathbf{x}$ and $\mathbf{x}'$ represent two corresponding image points taken from an undistorted IR/visible-light image pair. As a consequence the epipoles $\mathbf{e}$ and $\mathbf{e}'$, corresponding to the right and left null space of $\mathbf{F}$, are mapped to the point $\mathbf{p} = [1\ 0\ 0]^T$ at infinity. Since all epipolar lines must pass through their corresponding epipoles it is easy to verify that all epipolar lines run parallel to the x-axis and, in effect, all corresponding image points have identical $y$-coordinates. Fig. 6.12 shows the result of rectification for an arbitrary IR/visible-light image pair. Notice that due to the different field-of-views of the employed IR/visible-light camera pair, after rectification, the visible-light image is completely contained within the corresponding IR image. Moreover, Fig. 6.12 also illustrates the effect of distortion removal. This is particularly apparent when observing the boundaries of the IR image which, after

distortion removal, appear curved.

Upon completion of the rectification process, we manually displace the rectified images horizontally until the principal scene planes in the two views appear spatially aligned, crop the overlapping areas and resample the resulting image portions such that the final image resolution matches the native spatial resolution of the IR/visible-light video pair. The registration results of four IR/visible-light image pairs, recorded at different locations and with varying scene content, are depicted in Figs. 6.13 to 6.16.

Note that, in order for this displacement process to be automatic, a region of interest in the images, as well as corresponding points within it, would have to be identified. Such a region of interest would correspond to a given scene depth. Alternatively, given a set of corresponding salient points in the two images, the depths could be computed and a depth-based image rendering [157] could be used in one of the images to perform registration in all depths. However, such a method would not solve the problem of occluded areas between the two cameras.

## 6.5 Results

In order to show the effectiveness of the proposed IR/visible-light video registration framework, we performed experiments with 30 different video sequences, manually recorded at 6 distinct locations. Table 6.1 gives an overview of the main properties of the recorded video sequences, including a rough summary of the scene contents as well as the prevailing environmental conditions. Apart from the actual scene content, each IR/visible-light video pair starts off by exhibiting different poses of the calibration board of Fig. 6.4. These poses include translational and rotational movements of the calibration board and were chosen in such a way that both temporal and spatial alignment can be performed simultaneously using the same calibration footage.

The employed test setup, illustrated in Fig. 6.17, consisted of a portable tripod on which a pair of IR/visible-light cameras was rigidly mounted side-by-side. Moreover, the viewing angle and the zoom of the employed cameras were manually adjusted in such a way that the overlap between the field-of-view of both cameras was maximized.

The IR video sequences were obtained by recording the analogue NTSC video output of a FLIR Prism DS camera, operating at a spectral range of 3.6 to 5$\mu$m. In order to convert the analogue video stream to digital video, a Pinnacle Dazzle Digital Video Creator 150 video capturing device was utilized. In accordance with the NTSC standard, the resultant video exhibits a resolution of $720 \times 480$ pixels (which differs from the native resolution of the employed IR camera of $320 \times 244$ pixels). As for the

(a)

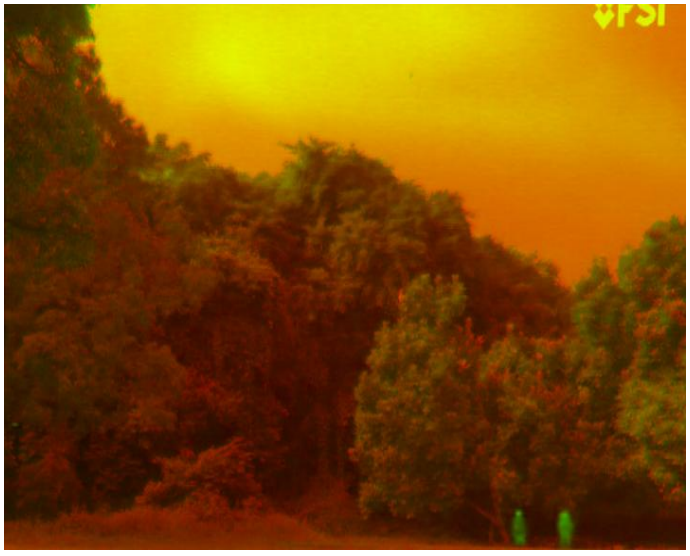
(b)

(c)

**Figure 6.13:** *Final registration results for an arbitrary IR/visible-light image pair from the "IPqM Baia 6" image sequence. (a) Registered visible-light image. (b) Registered IR image. (c) RGB pseudo-color image where the registered visible-light and IR images of (a) and (b) occupy the red and green color channels, respectively.*

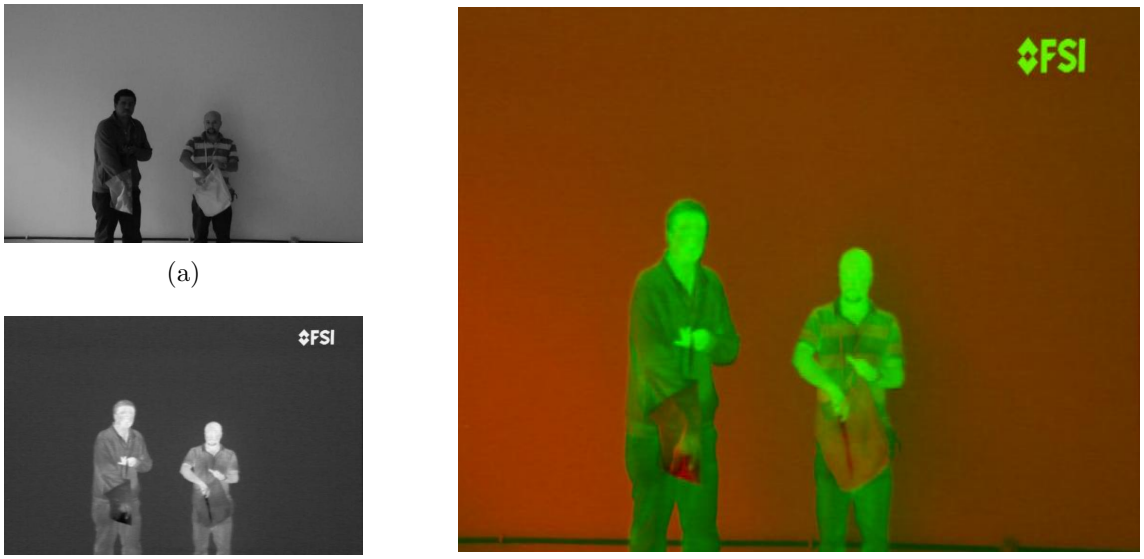
(a)

(b)



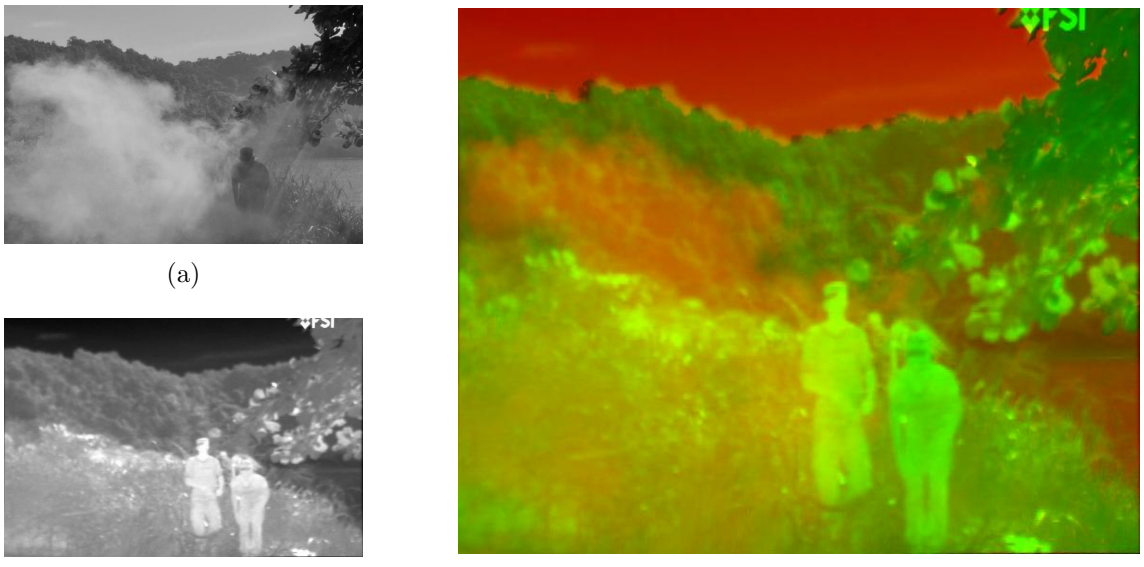
(c)

**Figure 6.14:** *Final registration results for an arbitrary IR/visible-light image pair from the "IPqM Campo 2" image sequence. (a) Registered visible-light image. (b) Registered IR image. (c) RGB pseudo-color image where the registered visible-light and IR images of (a) and (b) occupy the red and green color channels, respectively.*




(a) (b) (c)

**Figure 6.15:** *Final registration results for an arbitrary IR/visible-light image pair from the "IME Laboratório de Maquinas 1" image sequence. (a) Registered visible-light image. (b) Registered IR image. (c) RGB pseudo-color image where the registered visible-light and IR images of (a) and (b) occupy the red and green color channels, respectively.*



(a) (b) (c)

**Figure 6.16:** *Final registration results for an arbitrary IR/visible-light image pair from the "Forte São João 4" image sequence. (a) Registered visible-light image. (b) Registered IR image. (c) RGB pseudo-color image where the registered visible-light and IR images of (a) and (b) occupy the red and green color channels, respectively.*

| Name | Sequences | Scene Content | Environmental Conditions |
|---|---|---|---|
| Forte São João | 7 | Outdoor Scenes<br>People hiding behind vegetation and/or smoke screen<br>2 dominant scene planes at distances of approx. 10m and 300m, respectively | Bright sun light<br>30° Celcius |
| IME - Laboratório de Maquinas | 6 | Indoor Scenes<br>People walking around, hiding arms-like items within bags and behind newspapers<br>Distance to scene plane approx. 10m | Artificial light source<br>23° Celcius |
| IME - Lago | 2 | Outdoor Scenes<br>Several people passing by a corridor; One person hiding behind vegetation<br>Varying distance to scene plane (15m - 20m) | Twilight<br>28° Celcius |
| IPqM - Campo | 4 | Outdoor Scenes<br>2 people cross a lawn and hide behind trees; Crossing Car<br>Distance to scene plane approx. 50m | Bright sun light<br>33° Celcius |
| IPqM - Galpão | 5 | Indoor Scenes<br>Several people crossing a dimly lit corridor<br>Distance to scene plane approx. 15m | Artificial light source; Darkness<br>23° Celcius |
| IPqM - Baía | 6 | Outdoor Scenes<br>View of the Guanabara Bay and the bridge Rio de Janeiro - Niterói<br>Distance to scene plane approx. 500m | Nighttime<br>25° Celcius |

**Table 6.1:** *Overview of the recorded video sequences.*

visible-light video sequences a Panasonic HDC-TM700 camera was employed. The corresponding videos were recorded at a resolution of $1920 \times 1080$ and subsequently downsampled and cropped to match the IR video resolution of $720 \times 480$ pixels. Both IR and visible-light video sequences were recorded at a rate of 30 frames per second.

For the sake of brevity, we will only discuss the registration results for 6 different IR/visible-light video pairs, each recorded at a distinct location (see Table 6.1 for more details). However, since the calibration results are the same for all video pairs originating from the same location, the presented results can be considered valid for all recorded video sequences. Note that the same does not hold for the temporal alignment results which tend to differ from sequence to sequence. Representative scene thumbnails of the utilized IR/visible-light video sequences (before registration) are illustrated in Fig. 6.18.

**Figure 6.17:** *Utilized test setup consisting of an IR (left) and visible-light camera (right) mounted side-by-side.*

**Figure 6.18:** *Selected IR/visible-light scene thumbnails from all video sequences used for evaluation purposes. Top row consists of visible-light images, whereas the bottom row represents the corresponding IR images.*

### 6.5.1 Temporal Alignment Results

The estimated temporal offsets $\Delta\hat{t}$ for the 6 selected video sequences (see Fig. 6.18) together with the corresponding similarity measures of eq. (6.13) are given in Table 6.2. Note that the attained similarity is very close to one for all six assessed video sequences. This implies that after temporal alignment the movements of the calibration board are almost identical between the IR and visible-light video sequences. However, it is worth noting that the overall similarity measure depends, to a certain extent, on the performed movements with the calibration board. Thus, a lower similarity does not necessarily suggest a poor estimation of the temporal offset. Furthermore, Fig. 6.19 shows the obtained similarity measures over the whole set of temporal offset candidates for each assessed video pair. It can be observed that the curves always exhibit a single distinct peak at the position of the correct temporal offset, indicating the high robustness of the proposed framework.

Finally, in order to qualititavely demonstrate the effectiveness of the proposed temporal alignment scheme, Fig. 6.20 shows five calibration frames from the second IR/visible-light video sequence pair of Table 6.2 before and after temporal alignment. It can be noted that the unsynchronized video frames (Fig. 6.20(a)) display

| | 1st pair | 2nd pair | 3rd pair | 4th pair | 5th pair | 6th pair |
|---|---|---|---|---|---|---|
| Temporal Offset | 69 | 54 | 20 | 16 | 96 | 79 |
| Similarity | 0.9956 | 0.9951 | 0.9970 | 0.9989 | 0.9985 | 0.9974 |

**Table 6.2:** *Results of the temporal offset estimation for the six different IR/visible-light video sequence pairs corresponding to the scenes depicted in Fig. 6.18.*

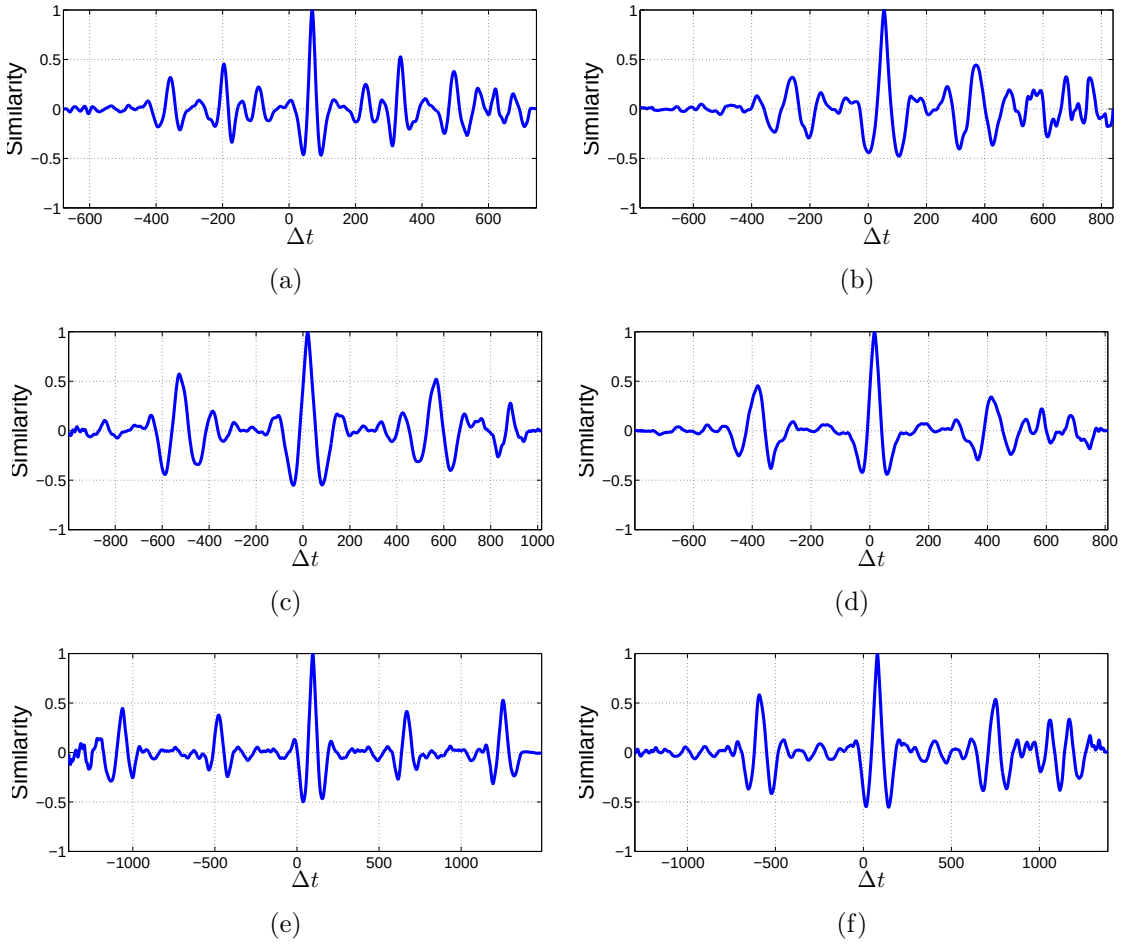
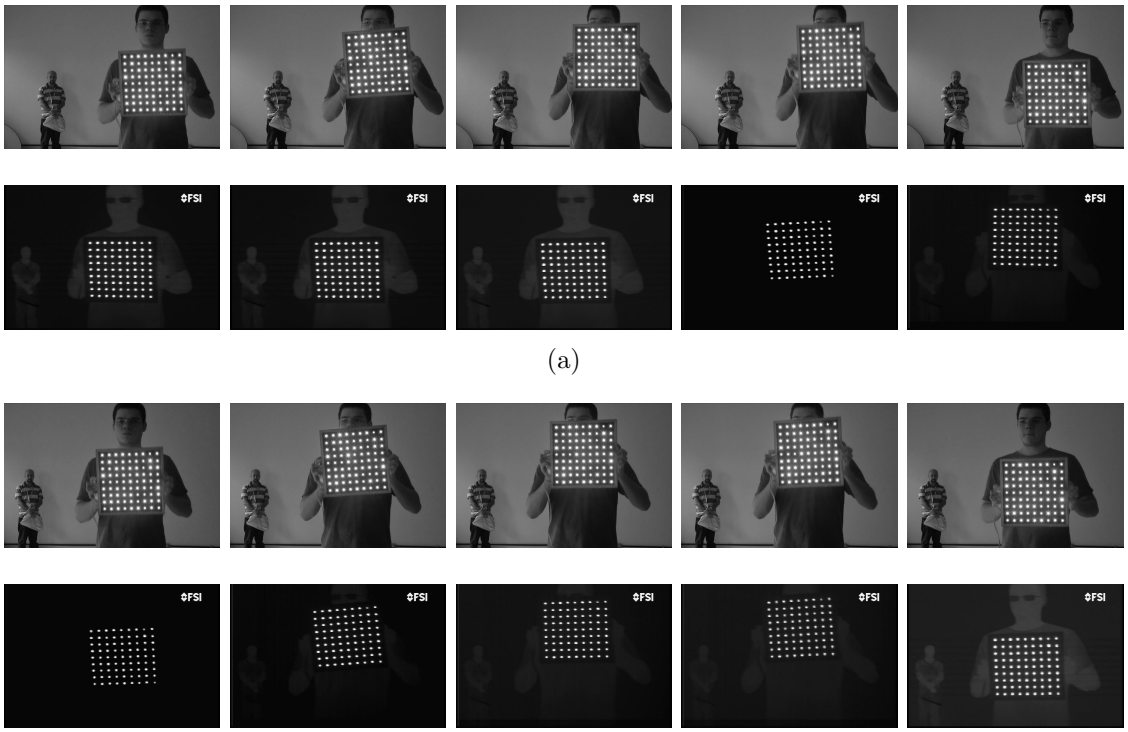



**Figure 6.19:** *Similarity measures over the whole set of possible temporal offset candidates corresponding to Table 6.2. (a) $1^{st}$ pair. (b) $2^{nd}$ pair. (c) $3^{rd}$ pair. (d) $4^{th}$ pair. (e) $5^{th}$ pair. (f) $6^{th}$ pair.*

a significant misalignment in time. This is particularly evident when observing the four IR video frames to the right which appear to lag considerably behind the visible-light frames. As for the synchronized video frames (Fig. 6.20(b)), both IR and visible-light frames exhibit similar poses of the alignment board, thus, indicating the correct temporal alignment of the IR/visible-light video sequence pair.

### 6.5.2 Calibration Results

After temporal alignment, the proposed calibration scheme was applied to 20 synchronized image pairs from each video sequence. The image pairs were chosen



(a)



(b)

**Figure 6.20:** *Five calibration frames of an arbitrary IR/visible-light video pair (a) before and (b) after temporal alignment.*

manually such that each frame exhibits a different pose of the calibration board of Fig. 6.4. As mentioned previously, the lens distortion of the employed cameras was assumed to comply with a $2^{\text{nd}}$ order radial distortion model with tangential distortion.

The obtained individual and stereo calibration results for the IR/visible-light camera pair corresponding to the "Forte São João" scenes are given in Table 6.3. Note that, for the sake of convenience, the rotation matrix $\mathbf{R}$ is given in the Rodrigues vector form $\mathbf{v}_{\text{rot}}$ [3]. Two things can be observed in Table 6.3. First of all, the focal length of the IR camera, measured in terms of pixels, considerably differs in the $x$ and $y$ direction. This suggests that the pixels are not square, consequently resulting in unequal scale factors along the $x$ and $y$ direction. Indeed, Fig. 6.21 indicates that the IR image appears to be stretched in the $x$-direction when compared to the visible-light image[1]. Note that the depicted images correspond to a synchronized IR/visible-light image pair before spatial registration.

Secondly, the principal point $(x_0, y_0)$ of the visible-light camera, which is expected to agree to a high degree with the center of the image, is noticeably off center. This is not an isolated case and was also observed in the calibration results of other video

[1]In fact, the focal length mismatch depicted in Table 6.3 suggests that the IR images do not get stretched along the $x$-axes but instead get contracted in the $y$-direction. We believe that this is most likely due to the aspect ratio change caused by the conversion to NTSC.

| Parameters | Visible-light Camera | | IR Camera | |
|---|---|---|---|---|
| | Estimated Values | $\sigma$ | Estimated Values | $\sigma$ |
| $\alpha_x$ | 1701.81 | 1.34 | 1785.54 | 1.72 |
| $\alpha_y$ | 1702.56 | 1.30 | 1595.44 | 1.49 |
| $x_0$ | 320.68 | 1.05 | 331.56 | 1.32 |
| $y_0$ | 182.96 | 1.12 | 259.04 | 1.26 |
| $s$ | 0.00 | 0.00 | 0.00 | 0.00 |
| $k_1$ | 0.02 | 0.00 | -0.45 | 0.00 |
| $p_1$ | -0.01 | 0.00 | 0.00 | 0.00 |
| $p_2$ | -0.02 | 0.00 | 0.01 | 0.00 |

(a)

| Parameters | Estimated Values | $\sigma$ |
|---|---|---|
| $\mathbf{v}_{\text{rot}}$ | $[-0.03\ 0.00\ 0.00]^T$ | $[0.00\ 0.00\ 0.00]^T$ |
| $\mathbf{t}$ | $[844.04\ 25.28\ 473.64]^T$ | $[10.96\ 10.15\ 3.95]^T$ |

(b)

**Table 6.3:** *(a) Individual and (b) Stereo camera calibration parameters corresponding to the IR/visible-light camera pair of the "Forte São João"video sequence. For the sake of convenience, the rotation matrix* $\mathbf{R}$ *is given in the Rodrigues vector form* $\mathbf{v}_{rot}$ *[3].*

sequences. We believe that this issue is closely related to the used calibration footage. In more detail, during recording we focused on capturing the calibration board in various rotational poses which, for the most part, are located in the center of the image. If instead we had moved the calibration board around the entire field-of-view of the camera, it could have been possible to turn the estimation of the principal point more robust. Nevertheless it is important to note that this assumption still awaits experimental validation.

In order to evaluate the accuracy of our calibration framework, we calculated the resulting mean reprojection error (MRE) when mapping the calibration point positions from the world coordinate system to the image plane, using the obtained set of calibration parameters. This was accomplished by computing the average MRE of eq. (6.10) over all 20 calibration images. Table 6.4 shows the resulting MREs for each video sequence together with the mean MRE obtained by averaging the MREs of the individual video sequences. At first glance it can be noted that the visible-light camera calibration results are consistently better than the calibration results for the IR camera. The reason for that is twofold. For starters, the point spread of the light bulbs in the visible-light calibration images is less accentuated than in the IR


**Figure 6.21:** *Stretching effect caused by the focal length mismatch between the IR and visible-light camera. The visible-light and IR images occupy the red and green channels, respectively, within the depicted RGB pseudo-color image.*

| Sequence Name | MRE | |
|---|---|---|
| | Visible-light Camera | IR Camera |
| Forte São João | 0.0288 | 0.0369 |
| IME - Laboratório de Maquinas | 0.0302 | 0.0349 |
| IME - Lago | 0.0235 | 0.0388 |
| IPqM - Campo | 0.0295 | 0.0381 |
| IPqM - Galpão | 0.0367 | 0.0442 |
| IPqM - Baía | 0.0233 | 0.0303 |
| **Average** | **0.0287** | **0.0372** |

**Table 6.4:** *MREs of the proposed IR/visible-light camera calibration method.*

calibration images (see Fig. 6.4). This allows for a more consistent calibration point detection along all 20 calibration images, consequently leading to better calibration results. Secondly, almost no lens distortion effects are present in the visible-light calibration images, further improving the overall calibration accuracy.

In order to assess the achieved results against the ones of the state-of-the-art, Table 6.5 lists the MREs for some selected IR/visible-light calibration schemes from the literature. Please note that these values were adapted via normalization, to match to the image resolution of each calibration image used in this work. This normalization was deemed necessary since, as reported in Table 6.5, different camera models with differing image resolutions were employed in the quoted references.

It can be noticed that our method appears to improve the calibration results almost by a factor of 2 for visible-light camera calibration and by a factor of 10 for IR camera calibration when compared to the state-of-the-art. However, it should be noted that due to possible differences in the simulation setup, a fair compari-

| Method | Camera | Resolution | MRE |
|---|---|---|---|
| Proposed | Panasonic HDC-TM700 | $720 \times 480$ | **0.0287** |
| Gschwandtner et al. [131] | Bumblebee XB3 | $1280 \times 960$ | 0.0475 |
| Vidas et al. [133] | Videre Apparen | $640 \times 480$ | 0.5151 |

(a)

| Method | Camera | Resolution | MRE |
|---|---|---|---|
| Proposed | FLIR Prism DS | $720 \times 480$ | **0.0372** |
| Yang et al. [130] | GUIDE IR112 | $320 \times 240$ | 1.2214 |
| Gschwandtner et al. [131] | Pathfind IR | $1280 \times 960$ | 0.4918 |
| Vidas et al. [133] | Miricle 307K | $640 \times 480$ | 0.3031 |

(b)

**Table 6.5:** *MREs of the proposed IR/visible-light camera calibration method and selected calibration schemes from the literature. (a) Visible-light camera calibration. (b) IR camera calibration. The MREs of the quoted references were adapted, via normalization, to match the image resolution of the calibration images used in this work.*

son cannot be conducted straightforwardly. Nevertheless, based on the vast differences between the MREs of the proposed scheme and the MREs of all remaining approaches, strong evidence exists that the proposed technique is indeed able to improve the accuracy of IR/visible-light camera calibration distinctly.

### 6.5.3 Image Fusion Example

As already pointed out in the introductory section of this chapter, the main motivation of this work was the creation of an image and video database suitable for image fusion purposes. Since such a collection would ideally include imagery of as many possible scenes as possible, recorded under a wide range of environmental conditions, we spent in total 3 months shooting 30 different videos at 6 different locations. The fusion scenarios were chosen in such a way that both visible-light and IR sequences convey complementary information about the scene such that, through fusion, a more complete picture of the scene can be achieved. Examples include the use of a smoke generator which, when turned on, generates a smoke screen that cannot be penetrated by the visible-light camera as well as the use of arms-like objects which appear solely in the IR images. Furthermore, we also recorded several nighttime scenes where IR imagery is essential to augment the overall scene information.

To this end, Fig. 6.22 shows the result of fusion for three selected IR/visible-light image pairs. As for the utilized fusion framework, we applied the proposed UWT fusion scheme with spectral factorization of Chapter 4 in conjunction with the 'Spline_3' filter bank of eq. (4.8) and four decomposition levels. The approximation

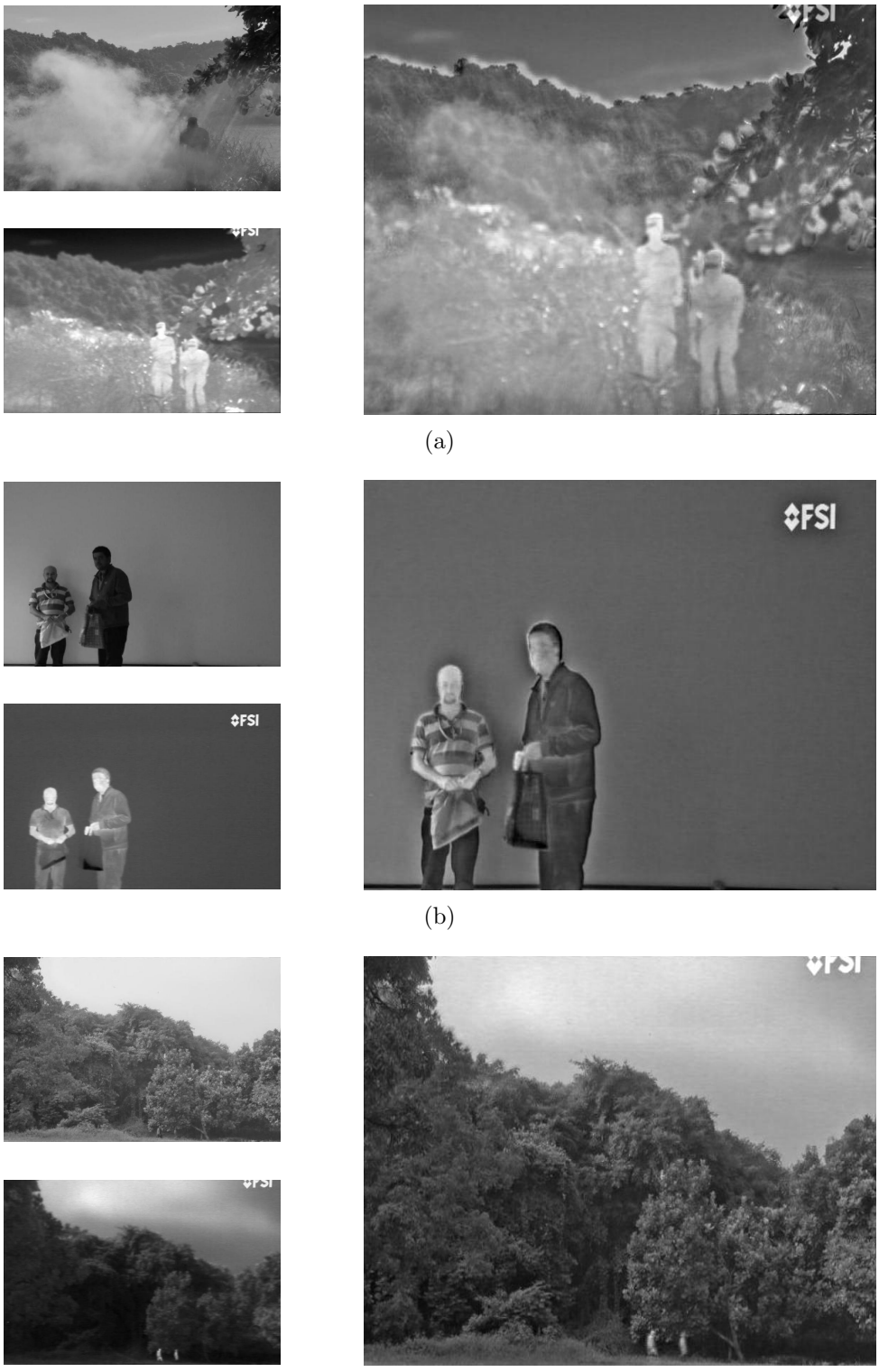



**Figure 6.22:** *Fusion results for selected IR/visible-light image pairs from the (a) "Forte São João 4", (b) "IME Laboratório de Maquinas 2" and (c) "IPqM Campo 2" video sequence. The fused images are depicted in the right column whereas the visible-light (top) and IR images (bottom) are located in the left column.*

images were fused using a simple averaging operation given in eq. (3.52) whereas the "choose max" fusion rule of eq. (3.51) was applied to the detail images.

By examining the obtained results it can indeed be noted that the fused images exhibit a more complete view of the overall scene. However, there also exists plenty of room for improvement. For example, the fused images appear to suffer from a considerable loss of contrast when compared to the source images. The reason for this is rooted in the use of the averaging fusion rule which tends to result in a destructive superposition when opposing approximation coefficients are added. This is particularly noticeable in image regions which are photographic negatives of each other, such as, for example, the sky in Fig. 6.22(a). Another interesting effect can be observed by looking at the transitional zone of forest and sky in the fused image of Fig. 6.22(a). Here, artificial flair was introduced into the fused image which is not present in any of the source images. The reason behind that is simple. Since the source images exhibit two principal scene planes with a large distance between each other, accurate pixel correspondences could only be accomplished for one of the planes. Thus, since registration was performed for the foreground plane, a noticeable pixel mismatch was introduced for the background plane which, during fusion, resulted in the creation of the depicted artifacts.

## 6.6 Conclusions

In this chapter a novel approach to IR/visible-light video registration has been introduced. Our method relies on a planar calibration board equipped with miniature light bulbs to increase the number of corresponding feature points within the frames of a temporally and spatially misaligned IR/visible-light video sequence pair. Thereby, the registration process is turned more robust against the chronic lack of mutual scene characteristics, which represent a common source of problems when registering video sequences originating from different spectral modalities.

The proposed processing chain first determines the exact light bulb positions in the individual frames of an IR/visible-light video sequence and utilizes this information to estimate the temporal offset. This is followed by the camera calibration process which is used to undistort and rectify the images such that the pixel coordinates in one image sequence are related to pixel coordinates in the other image sequence.

We showed in the course of this chapter that the proposed system is able to estimate the temporal offset with a very high confidence level. Furthermore, the introduced calibration scheme leads to calibration results which exhibit significantly smaller MREs when compared to the state-of-the-art. Finally, we demonstrated the effectiveness of the proposed framework for multimodal image fusion, where

co-registered images at sub-pixel accuracy are required.

In total 30 registered IR/visible-light video sequences, recorded at 6 different locations where generated in this work. They are available for download at `http://www.smt.ufrj.br/~fusion/`.

# Chapter 7

# Conclusions

This chapter concludes our investigations on multiscale image fusion. In what follows we will briefly summarize the main findings of the theoretic and practical work performed in the course of this project. The message conveyed in these conclusions represents more than four years of experience in the field of multiscale image fusion and completes the first part of our research efforts towards new and more powerful fusion algorithms.

Due to the vast popularity of multiscale fusion schemes, Chapter 2 started by giving a broad overview on the developments in this field of research. Special attention was drawn towards the question how multiscale decompositions can be meaningfully combined by the use of a varying set of fusion rules. Moreover, we presented a generic multiscale pixel-level framework which is able to encompass most of the existing multiscale fusion approaches found in the literature.

A large portion of the success of multiscale image fusion schemes depends on the choice of an appropriate transform. For this purpose a performance comparison of different multiscale transforms in the context of image fusion was conducted in Chapter 3. Based on the calculated objective fusion scores and the informal subjective assessment of the obtained fusion results, we concluded that the best fusion performance can be attained for redundant, shift-invariant transforms such as the Undecimated Wavelet Transform (UWT), the Dual-Tree Complex Wavelet Transform and the Nonsubsampled Contourlet Transform. Moreover, we observed that the overall behavior of multiscale fusion schemes considerably depends on the support size of the deployed filter bank - with a general tendency towards smaller filters.

The main contribution of this work was presented in Chapter 4. Based on the conclusions drawn in Chapter 3, we introduced a novel UWT-based pixel-level image fusion framework which splits the image decomposition process into two successive filtering operations using spectral factorization of the analysis filter pair. The actual fusion step takes place after convolution with the first filter pair, exhibiting a very

short support size. The underlying idea behind this approach is to minimize the undesirable spreading of coefficient values in the neighborhood of salient features whilst conserving the advantages of filters with higher orders. We showed that by following this strategy we are able to improve on the fusion results of other state-of-the-art fusion frameworks for a large group of input images, independent of the employed fusion rule. Another important feature of the presented approach was the use of non-orthogonal filter banks which are more robust to artifacts introduced during fusion compared to traditional (bi)orthogonal filter bank solutions.

A region-level extension of the fusion framework of Chapter 4 was proposed in Chapter 5. The basic idea here was to enhance the fusion results by including information about the presence of targets within the infrared image to the fusion process. For this purpose, we introduced a novel infrared-segmentation method which is able to detect targets in low-contrast environments whilst avoiding the introduction of spurious results. Additionally, we proposed a novel hybrid fusion scheme that utilizes both pixel- and region-level information to fuse infrared-visible source image pairs. The experimental results suggested that this methodology produces fused images with increased contrast and less artifacts around target regions. Finally, we demonstrated how target extraction can be used to artificially enhance the visibility of the extracted target regions.

Motivated by the lack of registered source images, Chapter 6 described the individual steps involved in the creation of an image data base for image fusion. For this purpose, a novel stereo camera calibration framework was introduced which is able to register a set of temporally and spatially misaligned IR/visible-light video sequences with very high precision. The proposed method utilized a planar calibration device equipped with miniature light bulbs to create a sufficient number of feature point correspondences between the input image pairs. Subsequently, these points were used to correct for the temporal offset and to spatially align the IR and visible-light video sequences. We showed that the proposed system is able to estimate the temporal offset with a very high confidence level and leads to calibration results exhibiting a significantly smaller mean reprojection error when compared to the state-of-the-art. Finally, we demonstrated the effectiveness of the proposed framework for multimodal image fusion. In the course of this work 30 registered IR/visible-light sequences were generated. They are available for download at `http://www.smt.ufrj.br/~fusion` and can be accessed freely by the research community to test and assess new image fusion schemes.

# Chapter 8

# Future work

Even though important insights could be gathered in the course of this work, the presented findings are still not exhaustive. Thus, in this chapter some natural extensions will be addressed.

## Artificial enhancement of the fusion results

As mentioned in various parts of this work, it is of vital importance that the fused image appears 'natural' and 'sharp' to a human interpreter. It seems therefore natural to attempt to improve the outcome of the fusion process by employing some sort of post-processing to the fused image. This should lead to a perceptually superior fused image when compared to the original result. For example, Bertalmío et al. [158] proposed a perceptual color correction technique which takes into account a set of human vision characteristics. Based on the fact that the human visual system is primarily sensitive to local contrast changes, the proposed scheme attempts to increase the contrast while respecting the visual content of the image (i.e., without introducing over-saturation or contouring effects).

Envisaged future work in this line of research will start with a thorough literature study on different image enhancement methods followed by an investigation of the possible implications of these techniques in the context of image fusion.

## Extension of the proposed fusion techniques to videos

The fusion frameworks of Chapter 4 and 5 utilize fusion rules which reach fusion decisions solely by considering information originating from a single source image pair. However, due to the availability of the image fusion data base of Chapter 6 containing a number of registered image sequences, a natural extension in this context would be to extend these frameworks to videos. In more detail, by introducing novel fusion rules incorporating information from adjacent frames into the

decision process it may be possible to consistently track objects-of-interest along the sequences and guarantee their inclusion into the fused frames.

Another area of future research lies within the question on how to guarantee the temporal stability/consistency of the fused sequences. Here, fusion rules need to be developed which take previous fusion decisions into account such that the fused sequence is free of abrupt contrast changes, among others, and appears natural to a human observer.

## Fusion guided image registration

As discussed in Chapter 6, upon completion of the rectification process, the rectified IR/visible-light image pairs need to be displaced manually until the principal scene planes in the two views appear aligned. However, this procedure is somehow unsatisfactory since it requires a certain degree of human interaction. One way to circumvent this problem is to identify a number of corresponding feature points within the scene plane and to dislocate the images until some similarity measure is maximized. However, as previously pointed out, such mutual feature points may not exist in the input images, consequently turning this approach impractical.

Another possible solution is to plug the outcome of image fusion into the image registration process. In more detail, for rectified IR/visible-light image pairs exhibiting a single scene plane, we may define the correct horizontal displacement as the one for which the objective fusion metrics of Section 3.2 reach their maximum, thus solving the correspondence problem.

For IR/visible-light image pairs exhibiting more than one principal scene plane, this process is considerably more challenging. A possible solution might be rooted in the use of the $Q_p$ fusion metric of eq. (3.48) which, as a preliminary step, provides us with two maps expressing the pixel-wise similarity between the first source image and the fused image as well as between the second source image and the fused image. When analyzing these maps for different horizontal offset candidates, it might be possible to identify local maxima, confined to some region within the map, for which the different scene planes are correctly aligned. However, for this process to be successful, knowledge of the exact number of scene planes must be available a-priori. Furthermore, such a method would still not be able to solve the problem of occluded areas between the two cameras.

# Bibliography


[1] HARTLEY, R., ZISSERMAN, A. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004.

[2] LI, S., YANG, B., HU, J. "Performance comparison of different multi-resolution transforms for image fusion", *Information Fusion*, v. 12, n. 2, pp. 74–84, 2011.

[3] FAUGERAS, O. *Three dimensional computer vision: A geometric viewpoint*. The MIT Press, 1993.

[4] PETROVIĆ, V. S. *Multisensor Pixel-level Image Fusion*. Ph.D Thesis, University of Manchester, Manchester, United Kingdom, 2001.

[5] ZHANG, Z., BLUM, R. S. "A categorization of multiscale-decomposition-based image fusion schemes with a performance study for a digital camera application", *Proceedings of the IEEE*, v. 87, n. 8, pp. 1315–1326, August 1999.

[6] ROCKINGER, O. "Pixel-level fusion of image sequences using wavelet frames". In: *Proceedings of the 16th Leeds Annual Statistical Research Workshop*, pp. 149–154, Leeds University Press, 1996.

[7] PIELLA, G. *Adaptive Wavelets and their Applications to Image Fusion and Compression*. Ph.D Thesis, University of Amsterdam, Amsterdam, Netherlands, 2003.

[8] LEWIS, J. J., O'CALLAGHAN, R. J., NIKOLOV, S. G., et al. "Pixel- and region-based image fusion with complex wavelets", *Information Fusion*, v. 8, n. 2, pp. 119–130, 2007.

[9] MITCHELL, H. B. *Image Fusion: Theories, Techniques and Applications*. 1 ed. Berlin, Springer-Verlag, 2010.

[10] ZHANG, Z., BLUM, R. S. "Region-Based Image Fusion Scheme For Concealed Weapon Detection". In: *Proceedings of the 31st Annual Conference on Information Sciences and Systems*, pp. 168–173, April 1997.

[11] VARSHNEY, P. K., CHEN, H.-M., RAMAC, L. C., et al. "Registration and fusion of infrared and millimeter wave images for concealed weapon detection". In: *Proceedings of the 1999 IEEE International Conference on Image Processing*, v. 3, pp. 532–536, 1999.

[12] YANG, J., BLUM, R. S. "A statistical signal processing approach to image fusion for concealed weapon detection". In: *Proceedings of the 2002 IEEE International Conference on Image Processing*, v. 1, pp. I–513 – I–516, 2002.

[13] CHEN, H.-M., LEE, S., RAO, R. M., et al. "Imaging for concealed weapon detection: a tutorial overview of development in imaging sensors and processing", *IEEE Signal Processing Magazine*, v. 22, n. 2, pp. 52–61, March 2005.

[14] BORGHYS, D., VERLINDE, P., PERNEEL, C., et al. "Multilevel data fusion for the detection of targets using multispectral image sequences", *Optical Engineering*, v. 37, n. 2, pp. 477–484, February 1998.

[15] GANG, X., BO, Y., ZHONGLIANG, X. "Infrared and visible dynamic image sequence fusion based on region target detection". In: *Proceedings of the 10th International Conference on Information Fusion*, pp. 772–776, 2007.

[16] DOBECK, G. J. "Fusing sonar images for mine detection and classification". In: *Proceedings of the SPIE*, v. 3710, pp. 602–614, 1999.

[17] TOET, A., IJSPEERT, J. K., WAXMAN, A. M., et al. "Fusion of visible and thermal imagery improves situational awareness", *Displays*, v. 18, n. 2, pp. 85–95, 1997.

[18] SCHOWENGERDT, R. A. *Remote Sensing: Models and Methods for Image Processing*. 3 ed. Burlington, Academic Press, 2007.

[19] TU, T.-M., SU, S.-C., SHYU, H.-C., et al. "A new look at IHS-like image fusion methods", *Information Fusion*, v. 2, n. 3, pp. 177–186, 2001.

[20] WANG, Z., ZIOU, D., ARMENAKIS, C., et al. "A comparative analysis of image fusion methods", *IEEE Transactions on Geoscience and Remote Sensing*, v. 43, n. 6, pp. 1391–1402, June 2005.

[21] KALPOMA, K. A., KUDOH, J.-I. "Image Fusion Processing for IKONOS 1-m Color Imagery", *IEEE Transactions on Geoscience and Remote Sensing*, v. 45, n. 10, pp. 3075 –3086, October 2007.

[22] NUNEZ, J., OTAZU, X., FORS, O., et al. "Multiresolution-based image fusion with additive wavelet decomposition", *IEEE Transactions on Geoscience and Remote Sensing*, v. 37, n. 3, pp. 1204–1211, May 1999.

[23] AIAZZI, B., ALPARONE, L., BARONTI, S., et al. "Context-driven fusion of high spatial and spectral resolution images based on oversampled multiresolution analysis", *IEEE Transactions on Geoscience and Remote Sensing*, v. 40, n. 10, pp. 2300–2312, October 2002.

[24] CHOI, M., KIM, R. Y., NAM, M.-R., et al. "Fusion of multispectral and panchromatic satellite images using the curvelet transform", *IEEE Geoscience and Remote Sensing Letters*, v. 2, n. 2, pp. 136–140, April 2005.

[25] NENCINI, F., GARZELLI, A., BARONTI, S., et al. "Remote sensing image fusion using the curvelet transform", *Information Fusion*, v. 8, n. 2, pp. 143–156, April 2007.

[26] DASARATHY, B. V. "Information fusion in the realm of medical applications - A bibliographic glimpse at its growing appeal", *Information Fusion*, v. 13, n. 1, pp. 1–9, January 2012.

[27] KOOY, H. M., VAN HERK, M., BARNES, P. D., et al. "Image fusion for stereotactic radiotherapy and radiosurgery treatment planning", *International Journal of Radiation Oncology*Biology*Physics*, v. 28, n. 5, pp. 1229–1234, 1994.

[28] KAM, M., ZHU, X., KALATA, P. "Sensor fusion for mobile robot navigation", *Proceedings of the IEEE*, v. 85, n. 1, pp. 108–119, January 1997.

[29] LEE, Y.-J., YIM, B.-D., SONG, J.-B. "Mobile Robot Localization based on Effective Combination of Vision and Range Sensors", *International Journal of Control Automation and Systems*, v. 7, n. 1, pp. 97–104, February 2009.

[30] SARI-SARRAF, H., GODDARD JR., J. S. "Vision system for on-loom fabric inspection", *IEEE Transactions on Industry Applications*, v. 35, n. 6, pp. 1252 –1259, November/December 1999.

[31] GROS, X. E., LIU, Z., TSUKADA, K., et al. "Experimenting with pixel-level NDT data fusion techniques", *IEEE Transactions on Instrumentation and Measurement*, v. 49, n. 5, pp. 1083–1090, October 2000.

[32] BURT, P. J., KOLCZYNSKI, R. J. "Enhanced image capture through fusion". In: *Proceedings of the 4th International Conference on Computer Vision*, pp. 173 –182, May 1993.

[33] GOSHTASBY, A. "Fusion of multi-exposure images", *Image and Vision Computing*, v. 23, n. 6, pp. 611–618, 2005.

[34] SHEN, R., CHENG, I., SHI, J., et al. "Generalized Random Walks for Fusion of Multi-Exposure Images", *IEEE Transactions on Image Processing*, v. 20, n. 12, pp. 3634–3646, December 2011.

[35] MITIANOUDIS, N., STATHAKI, T. "Pixel-based and region-based image fusion schemes using ICA bases", *Information Fusion*, v. 8, n. 2, pp. 131–142, 2007.

[36] CVEJIC, N., BULL, D., CANAGARAJAH, N. "Region-Based Multimodal Image Fusion Using ICA Bases", *IEEE Sensors Journal*, v. 7, n. 5, pp. 743–751, May 2007.

[37] BURT, P. J. "The pyramid as a structure for efficient computation". In: *Multiresolution Image Processing and Analysis*, Springer-Verlag, pp. 6–35, Berlin, 1984.

[38] LIU, Z., TSUKADA, K., HANASAKI, K., et al. "Image fusion by using steerable pyramid", *Pattern Recognition Letters*, v. 22, n. 9, pp. 929–939, 2001.

[39] PU, T., NI, G. Q. "Contrast-based image fusion using the discrete wavelet transform", *Optical Engineering*, v. 39, n. 8, pp. 2075–2082, August 2000.

[40] LI, H., MANJUNATH, B. S., MITRA, S. K. "Multisensor Image Fusion Using the Wavelet Transform", *Graphical Models and Image Processing*, v. 57, n. 3, pp. 235–245, 1995.

[41] PETROVIĆ, V. S., XYDEAS, C. S. "Gradient-based multiresolution image fusion", *IEEE Transactions on Image Processing*, v. 13, n. 2, pp. 228–237, February 2004.

[42] PAJARES, G., DE LA CRUZ, J. M. "A wavelet-based image fusion tutorial", *Pattern Recognition*, v. 37, n. 9, pp. 1855–1872, 2004.

[43] FORSTER, B., VAN DE VILLE, D., BERENT, J., et al. "Complex wavelets for extended depth-of-field: A new method for the fusion of multichannel microscopy images", *Microscopy Research and Technique*, v. 65, n. 1-2, pp. 33–42, September 2004.

[44] ROCKINGER, O. "Image sequence fusion using a shift-invariant wavelet transform". In: *Proceedings of the 1997 IEEE International Conference on Image Processing*, v. 3, pp. 288–291, October 1997.

[45] CHIBANI, Y., HOUACINE, A. "Redundant versus orthogonal wavelet decomposition for multisensor image fusion", *Pattern Recognition*, v. 36, n. 4, pp. 879–887, 2003.

[46] RAY, L. A., ADHAMI, R. R. "Dual tree discrete wavelet transform with application to image fusion". In: *Proceedings of the 38th Southeastern Symposium on System Theory*, pp. 430–433, March 2006.

[47] WAN, T., CANAGARAJAH, N., ACHIM, A. "Segmentation-Driven Image Fusion Based on Alpha-Stable Modeling of Wavelet Coefficients", *IEEE Transactions on Multimedia*, v. 11, n. 4, pp. 624–633, June 2009.

[48] LI, S., YANG, B. "Multifocus image fusion by combining curvelet and wavelet transform", *Pattern Recognition Letters*, v. 29, n. 9, pp. 1295–1301, July 2008.

[49] YANG, S., WANG, M., JIAO, L., et al. "Image fusion based on a new contourlet packet", *Information Fusion*, v. 11, n. 2, pp. 78–84, 2010.

[50] YANG, B., LI, S., SUN, F. "Image Fusion Using Nonsubsampled Contourlet Transform". In: *Proceedings of the 4th International Conference on Image and Graphics*, pp. 719–724, August 2007.

[51] ZHANG, Q., GUO, B.-L. "Multifocus image fusion using the nonsubsampled contourlet transform", *Signal Processing*, v. 89, n. 7, pp. 1334–1346, 2009.

[52] LI, S., YANG, B. "Hybrid Multiresolution Method for Multisensor Multimodal Image Fusion", *IEEE Sensors Journal*, v. 10, n. 9, September 2010.

[53] HUANG, N. E., SHEN, Z., LONG, S. R., et al. "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis", *Proceedings of the Royal Society of London Series A - Mathematical Physical and Engineering Sciences*, v. 454, n. 1971, pp. 903–995, March 1998.

[54] LOONEY, D., MANDIC, D. P. "Multiscale Image Fusion Using Complex Extensions of EMD", *IEEE Transactions on Signal Processing*, v. 57, n. 4, pp. 1626 –1630, April 2009.

[55] ZHANG, X. "Comparison of EMD Based Image Fusion Methods". In: *Proceedings of the 2009 International Conference on Computer and Automation Engineering*, pp. 302–305, 2009.

[56] SUN, Y.-Q., KOH, M. S., RODRIGUEZ-MAREK, E., et al. "A new infrared image fusion method using empirical mode decomposition and inpainting". In: *Proceedings of the 2011 IEEE International Conference on Image Processing*, pp. 1477–1480, September 2011.

[57] ROCKINGER, O., FECHNER, T. "Pixel-level image fusion: The case of image sequences". In: *Proceedings of the SPIE*, v. 3374, pp. 378–388, 1998.

[58] LALLIER, E., FAROOQ, M. "A real time pixel-level based image fusion via adaptive weight averaging". In: *Proceedings of the 3rd International Conference on Information Fusion*, v. 2, pp. WeC3/3–WeC3/13, July 2000.

[59] SHARMA, R. K., LEEN, T. K., PAVEL, M. "Probabilistic image sensor fusion". In: *Advances in Neural Information Processing Systems 11*, v. 11, pp. 824–830, 1999.

[60] KUMAR, M., DASS, S. "A Total Variation-Based Algorithm for Pixel-Level Image Fusion", *IEEE Transactions on Image Processing*, v. 18, n. 9, pp. 2137–2143, September 2009.

[61] ZRIBI, M. "Non-parametric and region-based image fusion with Bootstrap sampling", *Information Fusion*, v. 11, n. 2, pp. 85–94, April 2010.

[62] NEWMAN, E. A., HARTLINE, P. H. "The infrared vision of snakes", *Scientific American*, v. 246, n. 3, pp. 116–127, 1982.

[63] LI, M., CAI, W., TAN, Z. "A region-based multi-sensor image fusion scheme using pulse-coupled neural network", *Pattern Recognition Letters*, v. 27, n. 16, pp. 1948–1956, December 2006.

[64] HUANG, W., JING, Z. "Multi-focus image fusion using pulse coupled neural network", *Pattern Recognition Letters*, v. 28, n. 9, pp. 1123–1132, July 2007.

[65] WANG, Z., MA, Y., GU, J. "Multi-focus image fusion using PCNN", *Pattern Recognition*, v. 43, n. 6, pp. 2003–2016, 2010.

[66] WANG, Z., MA, Y. "Medical image fusion using m-PCNN", *Information Fusion*, v. 9, n. 2, pp. 176–185, April 2008.

[67] BROUSSARD, R. P., ROGERS, S. K., OXLEY, M. E., et al. "Physiologically motivated image fusion for object detection using a pulse coupled neural network", *IEEE Transactions on Neural Networks*, v. 10, n. 3, pp. 554–563, May 1999.

[68] ZITOVA, B., FLUSSER, J. "Image registration methods: a survey", *Image and Vision Computing*, v. 21, n. 11, pp. 977–1000, October 2003.

[69] TOMASI, C., KANADE, T. *Detection and Tracking of Point Features*. Relatório Técnico CMU-CS-91-132, Carnegie Mellon University, 1991.

[70] LOWE, D. "Distinctive Image Features from Scale-Invariant Keypoints", *International Journal of Computer Vision*, v. 60, n. 2, pp. 91–110, 2004.

[71] HARRIS, C., STEPHENS, M. "A combined corner and edge detector". In: *Proceedings of the 4th Alvey Vision Conference*, pp. 147–151, 1988.

[72] SZELISKI, R. "Image alignment and stitching: A tutorial", *Foundations and Trends® in Computer Graphics and Vision*, v. 2, n. 1, pp. 1–104, 2006.

[73] FISCHLER, M. A., BOLLES, R. C. "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography", *Communications of the ACM*, v. 24, n. 6, pp. 381–395, June 1981.

[74] DODGSON, N. A. "Quadratic interpolation for image resampling", *IEEE Transactions on Image Processing*, v. 6, n. 9, pp. 1322–1326, 1997.

[75] APPLEDORN, C. R. "A new approach to the interpolation of sampled data", *IEEE Transactions on Medical Imaging*, v. 15, n. 3, pp. 369–376, 1996.

[76] THÉVENAZ, P., BLU, T., UNSER, M. *Image interpolation and resampling, Handbook of Medical Image Processing*. New York, Academic Press, 2000.

[77] FIELD, D. J. "Scale-invariance and Self-similar 'Wavelet' Transforms: an Analysis of Natural Scenes and Mammalian Visual Systems". In: *Wavelets, Fractals and Fourier Transforms: New Developments and New Applications*, Oxford University Press, pp. 151–193, 1993.

[78] TOET, A. "Image fusion by a ratio of low-pass pyramid", *Pattern Recognition Letters*, v. 9, n. 4, pp. 245–253, 1989.

[79] KOREN, I., LAINE, A., TAYLOR, F. "Image fusion using steerable dyadic wavelet transform". In: *Proceedings of the 1995 IEEE International Conference on Image Processing*, v. 3, pp. 232–235, October 1995.

[80] PETROVIĆ, V. S., XYDEAS, C. S. "Cross band pixel selection in multiresolution image fusion". In: *Proceedings of the SPIE*, v. 3719, pp. 319–326, 1999.

[81] MALLAT, S. *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*. 3 ed. Burlington, Academic Press, 2009.

[82] DO, M. N., VETTERLI, M. "The Contourlet Transform: An Efficient Directional Multiresolution Image Representation", *IEEE Transactions on Image Processing*, v. 14, n. 12, pp. 2091–2106, December 2005.

[83] ESKICIOGLU, A. M., FISHER, P. S. "Image quality measures and their performance", *IEEE Transactions on Communications*, v. 43, n. 12, pp. 2959–2965, December 1995.

[84] DINIZ, P. S. R., DA SILVA, E. A. B., NETTO, S. L. *Digital Signal Processing: System Analysis and Design*. 2 ed. Cambridge, Cambridge University Press, 2010.

[85] DAUBECHIES, I. *Ten lectures on wavelets*. Society for Industrial and Applied Mathematics, 1992.

[86] VETTERLI, M., KOVAČEVIĆ. *Wavelets and Subband Coding*. Englewood Cliffs, Prentice-Hall, 1995.

[87] STRANG, G., NGUYEN, T. Q. *Wavelets and Filter Banks*. Wellesley, Wellesley Cambridge Press, 1996.

[88] ISO 15444-1:2004. *Information technology - JPEG-2000 image coding system: Core coding system*. Geneva, Switzerland, ISO.

[89] CANDES, E., DONOHO, D. "New tight frames of curvelets and optimal representations of objects with piecewise C-2 singularities", *Communications on Pure and Applied Mathematics*, v. 57, n. 2, pp. 219–266, February 2004.

[90] CANDES, E., DEMANET, L., DONOHO, D., et al. "Fast discrete curvelet transforms", *Multiscale Modeling & Simulation*, v. 5, n. 3, pp. 861–899, 2006.

[91] BURT, P., ADELSON, E. "The Laplacian Pyramid as a Compact Image Code", *IEEE Transactions on Communications*, v. 31, n. 4, pp. 532–540, 1983.

[92] DO, M. N. *Directional multiresolution image representations*. Tese de Doutorado, Swiss Federal Institute of Technology Lausanne, 2001.

[93] VETTERLI, M. "Multidimensional sub-band coding: Some theory and algorithms", *Signal Processing*, v. 6, n. 2, pp. 97–112, 1984.

[94] STARCK, J.-L., FADILI, J., MURTAGH, F. "The Undecimated Wavelet Decomposition and its Reconstruction", *IEEE Transactions on Image Processing*, v. 16, n. 2, pp. 297–309, February 2007.

[95] SHENSA, M. J. "The Discrete Wavelet Transform: Wedding the à Trous and Mallat Algorithms", *IEEE Transactions on Signal Processing*, v. 40, n. 10, pp. 2464–2482, October 1992.

[96] SELESNICK, I. W., BARANIUK, R. G., KINGSBURY, N. C. "The dual-tree complex wavelet transform", *IEEE Signal Processing Magazine*, v. 22, n. 6, pp. 123–151, November 2005.

[97] STARCK, J.-L., ELAD, M., DONOHO, D. "Redundant Multiscale Transforms and Their Application for Morphological Component Separation". v. 132, *Advances in Imaging and Electron Physics*, Elsevier, pp. 287–348, 2004.

[98] KINGSBURY, N. "Complex Wavelets for Shift Invariant Analysis and Filtering of Signals", *Applied and Computational Harmonic Analysis*, v. 10, n. 3, pp. 234–253, May 2001.

[99] DA CUNHA, A. L., ZHOU, J., DO, M. N. "The Nonsubsampled Contourlet Transform: Theory, Design, and Applications", *IEEE Transactions on Image Processing*, v. 15, n. 10, pp. 3089–3101, October 2006.

[100] PETROVIC, V. "Subjective tests for image fusion evaluation and objective metric validation", *Information Fusion*, v. 8, n. 2, pp. 208–216, April 2007.

[101] XYDEAS, C. S., PETROVIC, V. "Objective image fusion performance measure", *Electronics Letters*, v. 36, n. 4, pp. 308–309, February 2000.

[102] PIELLA, G., HEIJMANS, H. "A new quality metric for image fusion". In: *Proceedings of the 2003 IEEE International Conference on Image Processing*, v. 3, pp. III–173 – III–176, September 2003.

[103] QU, G., ZHANG, D., YAN, P. "Information measure for performance of image fusion", *Electronics Letters*, v. 38, n. 7, pp. 313–315, March 2002.

[104] LIU, Z., BLASCH, E., XUE, Z., et al. "Objective Assessment of Multiresolution Image Fusion Algorithms for Context Enhancement in Night Vision: A Comparative Study", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 34, n. 1, pp. 94 –109, January 2012.

[105] WANG, Z., BOVIK, A. C., SHEIKH, H. R., et al. "Image quality assessment: from error visibility to structural similarity", *IEEE Transactions on Image Processing*, v. 13, n. 4, pp. 600–612, April 2004.

[106] CVEJIC, N., CANAGARAJAH, C. N., BULL, D. R. "Image fusion metric based on mutual information and Tsallis entropy", *Electronics Letters*, v. 42, n. 11, pp. 626–627, May 2006.

[107] PHOONG, S.-M., KIM, C. W., VAIDYANATHAN, P. P., et al. "A new class of Two-Channel Biorthogonal Filter Banks and Wavelet Bases", *IEEE Transactions on Signal Processing*, v. 43, n. 3, pp. 649–665, March 1995.

[108] KINGSBURY, N. "A dual-tree complex wavelet transform with improved orthogonality and symmetry properties". In: *Proceedings of the 2000 IEEE International Conference on Image Processing*, v. 2, pp. 375–378, September 2000.

[109] ABDELNOUR, A. F., SELESNICK, I. W. "Design of 2-band orthogonal near-symmetric CQF". In: *Proceedings of the 2001 IEEE Conference on Acoustics, Speech, and Signal Processing*, v. 6, pp. 3693–3696, 2001.

[110] LU, W.-S., ANTONIOU, A., XU, H. "A direct method for the design of 2-D nonseparable filter banks". In: *Proceedings of the 1997 IEEE International Symposium on Circuits and Systems*, v. 4, pp. 2381–2384, June 1997.

[111] SHAH, A. I., KALKER, A. A. C. "Theory and design of multidimensional QMF sub-band filters from 1-D filters using transforms". In: *Proceedings of the 1992 International Conference on Image Processing and its Applications*, pp. 474–477, April 1992.

[112] CVETKOVIC, Z., VETTERLI, M. "Oversampled filter banks", *IEEE Transactions on Signal Processing*, v. 46, n. 5, pp. 1245–1255, May 1998.

[113] UNSER, M. "Ten Good Reasons For Using Spline Wavelets". In: *Proceedings of the 5th SPIE Conference on Wavelet Applications in Signal and Image Processing*, pp. 422–431, 1997.

[114] RODRIGUES, M. A. M. *Efficient Decompositions for Signal Coding*. Tese de Doutorado, COPPE/UFRJ, March 1999.

[115] NEVES, S. R., DA SILVA, E. A. B., MENDONCA, G. V. "Wavelet-watershed automatic infrared image segmentation method", *Electronics Letters*, v. 39, n. 12, pp. 903–904, June 2003.

[116] MEYER, F., BEUCHER, S. "Morphological segmentation", *Journal of Visual Communication and Image Representation*, v. 1, n. 1, pp. 21–46, 1990.

[117] MALLAT, S., ZHONG, S. "Characterization of signals from multiscale edges", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 14, n. 7, pp. 710–732, July 1992.

[118] BEUCHER, S., LANTUEJOUL, C. "Use of watersheds in contour detection". In: *Proceedings of the International Workshop on image processing, real-time edge and motion detection/estimation,*, 1979.

[119] PIELLA, G. "A region-based multiresolution image fusion algorithm". In: *Proceedings of the 5th International Conference on Information Fusion*, v. 2, pp. 1557–1564, 2002.

[120] TZAGKARAKIS, G., TSAKALIDES, P. "A statistical approach to texture image retrieval via alpha-stable modeling of wavelet decompositions". In: *Proceedings of the 5th Internation Workshop on Image Analysis for Multimedia Interactive Services*, April 2004.

[121] MA, X., NIKIAS, C. L. "Parameter estimation and blind channel identification in impulsive signal environments", *IEEE Transactions on Signal Processing*, v. 43, n. 12, pp. 2884–2897, December 1995.

[122] HAVIL, J. *Gamma: Exploring Euler's Constant*. Princeton, New Jersey, Princeton University Press, 2009.

[123] COVER, T. M., THOMAS, J. A. *Elements of Information Theory*. 2 ed. New York, Wiley, 2006.

[124] CASPI, Y., IRANI, M. "Spatio-temporal alignment of sequences", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 24, n. 11, pp. 1409–1424, 2002.

[125] HEIKKILÄ, J., SILVÉN, O. "A four-step camera calibration procedure with implicit image correction". In: *Proceedings of the 1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1106 – 1112, June 1997.

[126] HEIKKILÄ, J. "Geometric camera calibration using circular control points", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 22, n. 10, pp. 1066 – 1077, October 2000.

[127] ZHANG, Z. "A flexible new technique for camera calibration", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 22, n. 11, pp. 1330 – 1334, November 2000.

[128] DATTA, A., KIM, J.-S., KANADE, T. "Accurate camera calibration using iterative refinement of control points". In: *Proceedings of the 2009 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pp. 1201 – 1208, October 2009.

[129] PRAKASH, C., KARAM, L. "Camera calibration using adaptive segmentation and ellipse fitting for localizing control points". In: *Proceedings of the 2012 IEEE International Conference on Image Processing*, pp. 341–344, October 2012.

[130] YANG, R., YANG, W., CHEN, Y., et al. "Geometric Calibration of IR Camera Using Trinocular Vision", *Journal of Lightwave Technology*, v. 29, n. 24, pp. 3797 – 3803, December 2011.

[131] GSCHWANDTNER, M., KWITT, R., UHL, A., et al. "Infrared camera calibration for dense depth map construction". In: *Proceedings of the 2011 Intelligent Vehicles Symposium*, pp. 857 – 862, June 2011.

[132] LAGÜELA, S., GONZÁLEZ-JORGE, H., ARMESTO, J., et al. "Calibration and verification of thermographic cameras for geometric measurements", *Infrared Physics & Technology*, v. 54, n. 2, pp. 92 – 99, March 2011.

[133] VIDAS, S., LAKEMOND, R., DENMAN, S., et al. "A Mask-Based Approach for the Geometric Calibration of Thermal-Infrared Cameras", *IEEE Transactions on Instrumentation and Measurement*, v. 61, n. 6, pp. 1625–1635, 2012.

[134] STEIN, G. "Tracking from multiple view points: Self-calibration of space and time". In: *Proceedings of the 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, v. 1, pp. 521–527, 1999.

[135] LEE, L., ROMANO, R., STEIN, G. "Monitoring activities from multiple video streams: establishing a common coordinate frame", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 22, n. 8, pp. 758–767, 2000.

[136] CASPI, Y., IRANI, M. "Alignment of non-overlapping sequences". In: *Proceedings of the 2001 IEEE International Conference on Computer Vision*, v. 2, pp. 76–83, 2001.

[137] WOLF, L., ZOMET, A. "Sequence-to-Sequence Self Calibration". In: *Proceedings of the 2002 European Conference on Computer Vision*, v. 2351, pp. 370–382, 2002.

[138] WOLF, L., ZOMET, A. "Correspondence-free synchronization and reconstruction in a non-rigid scene". In: *Proceedings of the Workshop on Vision and Modelling of Dynamic Scenes*, May 2002.

[139] PADUA, F., CARCERONI, R., SANTOS, G., et al. "Linear Sequence-to-Sequence Alignment", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 32, n. 2, pp. 304–320, 2010.

[140] WEDGE, D., HUYNH, D., KOVESI, P. "Using Space-Time Interest Points for Video Sequence Synchronization". In: *Proceedings of the IAPR Conference on Machine Vision Applications*, pp. 190–194, 2007.

[141] CASPI, Y., SIMAKOV, D., IRANI, M. "Feature-Based Sequence-to-Sequence Matching", *International Journal of Computer Vision*, v. 68, n. 1, pp. 53–64, 2006.

[142] DAI, C., ZHENG, Y., LI, X. "Accurate Video Alignment Using Phase Correlation", *IEEE Signal Processing Letters*, v. 13, n. 12, pp. 737–740, 2006.

[143] RAVICHANDRAN, A., VIDAL, R. "Video Registration Using Dynamic Textures", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 33, n. 1, pp. 158–171, 2011.

[144] CASPI, Y., IRANI, M. "A step towards sequence-to-sequence alignment". In: *Proceedings of the 2000 IEEE Conference on Computer Vision and Pattern Recognition*, v. 2, pp. 682–689, 2000.

[145] UKRAINITZ, Y., IRANI, M. "Aligning Sequences and Actions by Maximizing Space-Time Correlations". In: *Proceedings of the 2006 European Conference on Computer Vision*, v. 3953, pp. 538–550, 2006.

[146] USHIZAKI, M., OKATANI, T., DEGUCHI, K. "Video Synchronization Based on Co-occurrence of Appearance Changes in Video Sequences". In: *Proceedings of the 2006 International Conference on Pattern Recognition*, v. 3, pp. 71–74, 2006.

[147] SAWHNEY, H., KUMAR, R. "True multi-image alignment and its application to mosaicing and lens distortion correction", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 21, n. 3, pp. 235–243, 1999.

[148] PRAKASH, S., LEE, P., CAELLI, T., et al. "Robust thermal camera calibration and 3D mapping of object surface temperatures". In: *Proceedings of the XXVIII SPIE Conference on Thermosense*, v. 6205, 2006.

[149] BRADSKI, G., KAEHLER, A., PISAREVSKY, V. "Learning-based computer vision with intel's open source computer vision library", *Intel Technology Journal*, v. 9, n. 2, pp. 119 – 130, May 2005.

[150] BOUGUET, J.-Y. "Camera Calibration Toolbox for Matlab". `http://www.vision.caltech.edu/bouguetj/calib_doc/`, 2012. [Online; accessed 12/12/2012].

[151] CANNY, J. "A Computational Approach to Edge Detection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 8, n. 6, pp. 679 – 698, November 1986.

[152] BALLARD, D. "Generalizing the Hough transform to detect arbitrary shapes", *Pattern Recognition*, v. 13, n. 2, pp. 111–122, 1981.

[153] FITZGIBBON, A., PILU, M., FISHER, R. "Direct least-squares fitting of ellipses", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 21, n. 5, pp. 476–480, May 1999.

[154] FAUGERAS, O., LUSTMAN, F. "Motion and Structure from Motion in a piecewise planar environment", *International Journal of Pattern Recognition and Artificial Intelligence*, v. 02, n. 03, pp. 485–508, 1988.

[155] ZHANG, Z., HANSON, A. "3D reconstruction based on homography mapping". In: *ARPA Image Understanding Workshop*, pp. 249–399, 1996.

[156] LOOP, C., ZHANG, Z. "Computing rectifying homographies for stereo vision". In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, v. 1, 1999.

[157] DEBEVEC, P. E., TAYLOR, C. J., MALIK, J. "Modeling and rendering architecture from photographs: A hybrid geometry-and image-based approach". In: *Proceedings of the 23rd annual conference on Computer Graphics and Interactive Techniques*, pp. 11–20, 1996.

[158] BERTALMIO, M., CASELLES, V., PROVENZI, E., et al. "Perceptual Color Correction Through Variational Techniques", *IEEE Transactions on Image Processing*, v. 16, n. 4, pp. 1058 –1072, April 2007.

# List of publications


[1] DA SILVA, E. S., PAGLIARI, C. L., DA SILVA, E. A. B., ELLMAUTHALER, A. "Análise Comparativa de Métodos de Fusão de Imagens". In: *Proceedings of the XXIX Simpósio Brasileiro de Telecomunicações*, Curitiba, Brazil, October 2011.

[2] ELLMAUTHALER, A., DA SILVA, E. A. B., PAGLIARI, C. L., NEVES, S. R. "Infrared-Visible Image Fusion using the Undecimated Wavelet Transform with Spectral Factorization and Target Extraction". In: *Proceedings of the 2012 International Conference on Image Processing*, Orlando, September 2012.

[3] ELLMAUTHALER, A., DA SILVA, E. A. B., PAGLIARI, C. L., PEREZ, M. M. "Multiscale Image Fusion Using the Undecimated Wavelet Transform With Non-Orthogonal Filter Banks". In: *Proceedings of the XXX Simpósio Brasileiro de Telecomunicações*, Brasília, Brazil, September 2012.

[4] ELLMAUTHALER, A., PAGLIARI, C. L., DA SILVA, E. A. B. "Multiscale Image Fusion Using the Undecimated Wavelet Transform With Spectral Factorization and Non-Orthogonal Filter Banks". *IEEE Transactions on Image Processing*, v. 22, n. 3, pp. 1005-1017, March 2013.

[5] ELLMAUTHALER, A., DA SILVA, E. A. B., PAGLIARI, C. L., GOIS, J. N. "A Robust Temporal Alignment Technique for Infrared and Visible-Light Video Sequences". In: *Proceedings of the XXXI Simpósio Brasileiro de Telecomunicações*, Fortaleza, Brazil, September 2013.

[6] ELLMAUTHALER, A., DA SILVA, E. A. B., PAGLIARI, C. L., NEVES, S. R., GOIS, J. N. "A Novel Iterative Calibration Approach for Thermal Infrared Cameras". In: *Proceedings of the 2013 International Conference on Image Processing*, Melbourne, September 2013.