Optimization of the compression/restoration chain for satellite images

Mikaël Carlavan

UNIVERSITY OF NICE - SOPHIA ANTIPOLIS

DOCTORAL SCHOOL STIC

SCIENCES ET TECHNOLOGIES DE L'INFORMATION
ET DE LA COMMUNICATION

P H D T H E S I S

to obtain the title of

PhD of Science

of the University of Nice - Sophia Antipolis

Speciality : Computer Science

Prepared by

Mikael Carlavan

Optimization of the compression/restoration chain for satellite images

Supervised by Laure Blanc-Féraud and Marc Antonini

and prepared at I3S Sophia Antipolis in the Morpheme and MediaCoding teams

Jury :

President :	Claude Labit	-	Director of Research (INRIA)
Reviewers :	Andres Almansa	-	Research Scientist (Telecom ParisTech)
	Philippe Salembier	-	Professor (UPC)
Advisors :	Laure Blanc-Féraud	-	Director of Research (I3S)
	Marc Antonini	-	Director of Research (I3S)
Members :	Caroline Chaux	-	Research Scientist (LATP)
	Carole Thiebaut	-	Engineer (CNES)
	Yves Bobichon	-	Engineer (Thales Alenia Space)

Contents

List of Abbreviations

CCSDS Consultative Committee for Space Data Systems

CDF Cohen-Daubechies-Feauveau

CS Compressed sensing

DCT Discrete cosine transform

EBCOT Embedded block coding with optimized truncation

EZW Embedded zerotree wavelet

GGD Generalized Gaussian distribution

KKT Karush-Kuhn-Tucker

KLT Karhunen-Loève transform

LSB Least significant bit

MSB Most significant bit

MSE Mean square error

MTF Modulation transfer function

OTF Optical transfer function

PSF Point spread function

PSNR Peak signal-to-noise ratio

SNR Signal-to-noise ratio

SPIHT Set partitioning in hierarchical trees

TV Total variation

Abstract

The subject of this work is image coding and restoration in the context of satellite imaging. Regardless of recent developments in image restoration techniques and embedded compression algorithms, the reconstructed image still suffers from coding artifacts making its quality evaluation difficult. The objective of the thesis is to improve the quality of the final image with the study of the optimal structure of decoding and restoration regarding to the properties of the acquisition and compression processes. More essentially, the aim of this work is to propose a reliable technique to address the optimal decoding-deconvolution-denoising problem in the objective of global optimization of the compression/restoration chain.

The thesis is organized in three parts. The first part is a general introduction to the problematic addressed in this work. We then review a state-of-the-art of restoration and compression techniques for satellite imaging and we describe the current imaging chain used by the French Space Agency (CNES[1]) as this is the focus of the thesis.

The second part is concerned with the global optimization of the satellite imaging chain. We propose an approach to estimate the *theoretical* distortion of the complete chain and we present, for three different configurations of coding/restoration, an algorithm to perform its minimization. Our second contribution is also focused on the study of the global chain but is more aimed to optimize the visual quality of the final image. We present *numerical* methods to improve the quality of the reconstructed image and we propose a novel imaging chain based on the image quality assessment results of these techniques.

The last part of the thesis introduces a satellite imaging chain based on a new sampling approach. This approach is interesting in the context of satellite imaging as it allows to transfer all the difficulties to the on-ground decoder. We recall the main theoretical results of this sampling technique and we present a satellite imaging chain based on this framework. We propose an algorithm to solve the reconstruction problem and we conclude by comparing the proposed chain to the one currently used by the CNES.

[1] Centre National d'Etudes Spatiales

Résumé

Le sujet de cette thèse concerne le codage et la restauration d'image dans le contexte de l'imagerie satellite. En dépit des récents développements en restauration et compression embarquée d'images, de nombreux artéfacts apparaissent dans la reconstruction de l'image. L'objectif de cette thèse est d'améliorer la qualité de l'image finale en étudiant la structure optimale de décodage et de restauration en fonction des caractéristiques des processus d'acquisition et de compression. Plus globalement, le but de cette thèse est de proposer une méthode efficace permettant de résoudre le problème de décodage-déconvolution-débruitage optimal dans un objectif d'optimisation globale de la chaîne compression/restauration.

Le manuscrit est organisé en trois parties. La première partie est une introduction générale à la problématique traitée dans ce travail. Nous présentons un état de l'art des techniques de restauration et de compression pour l'imagerie satellite et nous décrivons la chaîne de traitement actuellement utilisée par le Centre National d'Etudes Spatiales (CNES) qui servira de référence tout au long de ce manuscrit.

La deuxième partie concerne l'optimisation globale de la chaîne d'imagerie satellite. Nous proposons une approche pour estimer la distorsion *théorique* de la chaîne complète et développons, dans trois configurations différentes de codage/restauration, un algorithme pour réaliser la minimisation. Notre deuxième contribution met également l'accent sur l'étude la chaîne globale mais est plus ciblée sur l'optimisation de la qualité visuelle de l'image finale. Nous présentons des méthodes *numériques* permettant d'améliorer la qualité de l'image reconstruite et nous proposons une nouvelle chaîne image basée sur les résultats d'évaluation de qualité de ces techniques.

La dernière partie de la thèse introduit une chaîne d'imagerie satellite basée sur une nouvelle théorie de l'échantillonnage. Cette technique d'échantillonnage est intéressante dans le domaine du satellitaire car elle permet de transférer toutes les difficultés au décodeur qui se situe au sol. Nous rappelons les principaux résultats théoriques de cette technique d'échantillonnage et nous présentons une chaîne image construite à partir de cette méthode. Nous proposons un algorithme permettant de résoudre le problème de reconstruction et nous concluons cette partie en comparant les résultats obtenus avec cette chaîne et celle utilisée actuellement par le CNES.

Part I

Introduction

Chapter 1

Presentation of the thesis

1.1 Context and motivations

Satellite imaging has been the focus of intense works in the remote sensing community for the last years. The ability of satellite optical systems to produce high resolution images has indeed been of a great interest in applications such as change detection or image classification. It has however outcomed to be quite challenging for the design of satellite imaging chains. The dimension of images acquired by high-resolution satellites keeps growing as the image resolution, i.e. the spatial distance between two adjacent pixels, gets smaller while the swath maintains. For example, one image of the PLEIADES-HR satellite covers an area of 20 km $\times$ 20 km with a resolution of 70 cm, giving an image size of almost 30000×30000 pixels. These images are quantized on 12 bits, which represents 1.35 Gb of raw data per image! In addition, a satellite is not able to continuously transmit the acquired images as ground stations are not always accessible for a transmission. It has to store the acquired images on the on-board mass storage to transmit them later. But the on-board storage capacity of a satellite is highly limited (about 500 Gb for the PLEIADES-HR satellite [Lier 2008]) such that the on-board memory needs to be cleared frequently; the step of image coding is then important and stands as a major element of the satellite imaging chain. The step of restoration is also very important. Due to the constraint on the size of the optics, the acquired image is blurred and a deconvolution/denoising process is always required to produce an image which can be exploited.

Despite the recent advances in image coding, many artifacts appear on the reconstructed image. These artifacts appear as specific patterns which clearly interfere with the image quality assessment. In this sense, the objective of the thesis is to improve the quality of the final image with the study of the optimal decoding structure regarding to the characteristics of the acquisition and compression chains. More generally, the aim of this work is to bring a methodological contribution to the optimal decoding-deconvolution-denoising problem and consists in a characterization and an optimization of the compression/restoration chain considering the instrumental characteristics. As part of the thesis, we do not constrain the complexity of proposed on-board algorithms et we assume that future electronics architectures will allow to embed these algorithms. Works on this subject are currently in progress at the French Space Agency (CNES).

To formulate this specific global optimization problem, we consider the imaging chain showed Fig. 1.1. We denote by x the analog scene. Depending on the context,

x may also be referred in the thesis to the reference or target image which is the closest discrete representation of the true analog scene that we can obtain (we will detail this aspect in Chapter 3). The acquired image y is the image collected after the sampling and the analog-to-digital conversion. This image is the direct ouput of the optical instrument and therefore will be referred as the instrumental image. This image is encoded on-board of the satellite to form a compressed bitstream such that it can be efficiently stored then transmitted to the ground station. The complexity of the coding scheme is strongly constrained by the resources available on board which remain highly limited, such that the design of this step is usually a difficult task. The evolution of electronics parts and on-board satellite architectures may however allow more complex algorithms for future missions.

Once the encoded image has been transmitted to the ground station, it is decoded and a restoration is applied to reduce the degradations due to the acquisition and the coding processes. The restored image is the final image and is denoted $\hat{x}$. This image is the image obtained after the coding/decoding C and the restoration T and should be the closest representation (following some distance that we will define) of the reference image.

We denote by $D(x, \hat{x})$ some measure of the distance between the reference image and the restored one. In the considered chain, the coded/decoded image is $C(y)$ (the decoding operator is included in C for more clarity in the notations) and we will denote $R(C(y))$ some measure of the coding rate of the coded image. The restored image is obtained by applying the restoration operator T on the coded/decoded image $C(y)$. It can then be expressed as a function of the coding and the restoration by $\hat{x} = T(C(y))$. The problem of global optimization consists in finding the optimal C^* and T^* which minimize the distance $D(x, \hat{x})$ under the constraint that the target rate R_c is not exceeded. This can be formulated as

$$C^*, T^* = \begin{array}{ll} \arg\min & E\left[D(x, T(C(y)))\right] \\ \text{subject to} & C, T \\ & R(C(y)) \leq R_c \end{array}, \tag{1.1}$$

where E is the expected value with respect to the distribution law of x, meaning that we want to minimize on average the distance $D(x, \hat{x})$ for all images x which follow a certain probability distribution.

Solving problem (1.1) is very difficult in many aspects. Firstly, problem (1.1) searches for the optimal coder and restoration among all techniques, which is not tractable. Second, even if the coding and restoration methods are given and perfectly known, an analytic expression of the global distortion is usually not available as the coder and the restoration are highly complex and can rarely be expressed in closed-form. Moreover, the global distortion depends on the knowledge of the real unknown image x (or its statistics) and on the distance measure D. Ideally, D should evaluate the image quality with the same accuracy as image analysis experts. Designing such criterion is however difficult and out of the topic of the thesis. In this work, we will always take D to be equal to the mean square error since it is a tool that we can easily manipulate. We are aware that the mean square error is not the best criterion

that we can use, we will see however that its flexibility is very interesting to develop global optimization techniques. But as we can see, the problem (1.1) is difficult to solve in a general context.

The contribution of the thesis is then to bring some insights on the global optimization of the imaging chain. We will first focus on the theoretical optimization of the global distortion in the case of a simple imaging chain. Even if the considered chain is overly simple, the proposed method appears to be original and tackles a major difficulty in formulating a closed-form expression of the global distortion. Because of the complexity of a true satellite imaging chain, we will then present several experiments to optimize the quality of the final image. This numerical study addresses common questions in the design of the imaging chain such as the position of the restoration (i.e. on-board before coding or on-ground after decoding) and how to process the coding artifacts which interfere with the interpretation of the image. To conclude the thesis, we will study a new imaging chain based on recent advances in the theory of sampling. This theory appears at first slightly opposing the current imaging chain. But the benefit in term of embedded resources clearly justify our interest to this method.

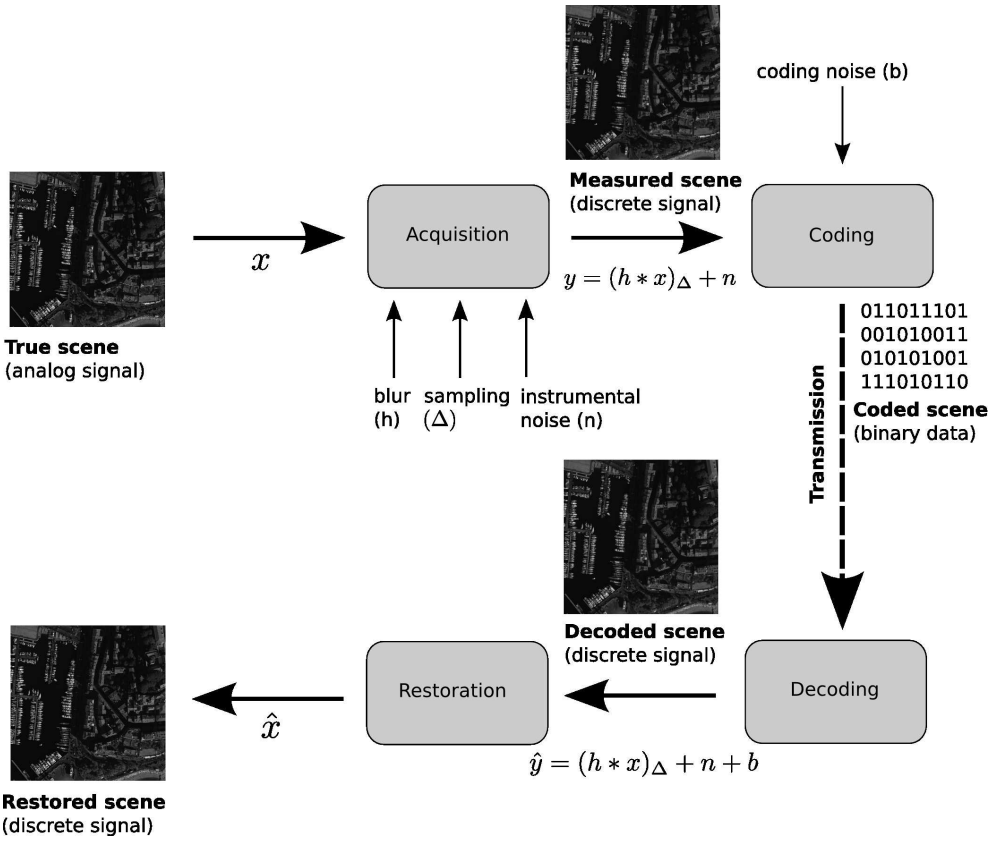


Figure 1.1: Current processing chain for satellite imaging.

1.2 Organization of the thesis

This document is divided in three parts. Part I is a general introduction to the thesis. In this part, Chapter 1 describes the context and the organization of the manuscript. The chapter 2 presents a state-of-the-art of restoration and compression techniques for satellite images. Chapter 3 closes the part by the technical description of the current imaging chain used by the CNES, which is the focus of the thesis.

Part II is the core of the thesis and is concerned with the global optimization of the satellite imaging chain. This study is our main contribution and is divided in two chapters. Chapter 4 is focussed on the theoretical optimization of the chain. In this chapter, we consider a simple case of imaging chain and we propose a model to estimate the global distortion. This estimation is then minimized with respect to the parameters of the chain to get the minimum global distortion (and the optimal parameters) given a target coding rate. The main result of this optimization is that the quality of the final image can be highly improved if we address the problem of the satellite imaging chain optimization in its globality. This chapter also addresses theoretically the question of the position of the restoration in the imaging chain.

The other part of our work is described in Chapter 5 and is also focused on the optimization of the chain but the true satellite imaging chain is now considered. Due to the difficulty to extend the previous study to this chain, we present in Chapter 5 a set of numerical experiments which improve the quality of the final image. Through this experimental study, Chapter 5 addresses recurrent open questions such as the position of the restoration in the chain and how to deal with the coding noise. From the obtained results, we propose a new satellite imaging chain based on an on-board restoration coupled with a subtractive dithering technique. Compared to the current imaging chain, the proposed approach eliminates several current problems in the observation of the final image such as structured coding artifacts.

Finally Part III introduces a satellite imaging chain based on the compressed sensing approach. In Chapter 6, we recall the main results of the compressed sensing theory and we present a satellite imaging chain based on this framework. We propose an algorithm to solve the reconstruction problem and we conclude by comparing the proposed chain to the current imaging chain.

1.3 Publications

Journal papers

- M. Carlavan, L. Blanc-Féraud, M. Antonini, C. Thiebaut, C. Latry and Y. Bobichon. Joint coding-denoising optimization of noisy images. *Submitted to IEEE Transactions on Image Processing.*

- M. Carlavan, L. Blanc-Féraud, M. Antonini, C. Thiebaut, C. Latry and Y. Bobichon. On the optimization of the satellite imaging chain. *Submitted to IEEE Transactions on Geoscience and Remote Sensing.*

International conferences

- M. Carlavan, L. Blanc-Féraud, M. Antonini, C. Thiebaut, C. Latry and Y. Bobichon. Global rate-distortion optimization of satellite imaging chains. *On-Board Payload Data Compression Workshop (OPBDC)*, Oct. 2012.

- M. Carlavan, L. Blanc-Féraud, M. Antonini, C. Thiebaut, C. Latry and Y. Bobichon. A satellite imaging chain based on the Compressed Sensing technique. *On-Board Payload Data Compression Workshop (OPBDC)*, Oct. 2012.

National conferences

- M. Carlavan, L. Blanc-Féraud, M. Antonini, C. Thiebaut, C. Latry and Y. Bobichon. Optimisation jointe de la chaîne codage/débruitage pour les images satellite. *Submitted to GRETSI*, 2013.

Chapter 2

State-of-the-art of optimization techniques for satellite imaging

In this chapter, we make a brief review of optimization techniques applied to the satellite imaging chain. We distinct here two types of optimization techniques:

- The techniques which optimize only one component of the chain regardless to the other ones. This type of optimization is referred in this thesis as separate or disjoint optimization.
- The techniques which optimize one component of the chain by taking into account the characterics of the other ones. This type of optimization is referred in this thesis as joint optimization.

We organized this chapter in two sections and we discuss each type of optimization technique in each section. Section 2.1 starts this review by presenting advanced coding and restoration techniques. Although the mentioned techniques have not been specifically designed for satellite imaging, they are often used as basis in the design of these parts. Section 2.2 is dedicated to coding and restoration techniques designed to globally optimize the satellite imaging chain. In this part, we present the methods proposed in [Parisot 2000a] and in [Tramini 1998] which are, to the best of our knowledge, the two main existing contributions in this domain.

2.1 Disjoint optimization techniques

2.1.1 Advanced compression algorithms

The information inside an image (and more specifically in a high resolution one) is strongly redundant (refer, for example, to the image of Cannes harbour Fig. 2.2). It is then possible to compress a satellite image by reducing this redundancy without losing important features. It is indeed unusual that the totality of an image brings relevant information and one can reach significant compression rates if one accepts to slightly deteriorate its quality. This is the process of lossy compression. Such a compression technique is composed of several steps as shown on the Fig. 2.1.

The first step of a lossy compression scheme is to decorrelate the data. The idea of the decorrelation step is to reduce the redundancy in an image by using a (most of time linear) transform which gathers all its energy in a small number of non-null coefficients, usually located in the low frequencies of the signal. These transforms

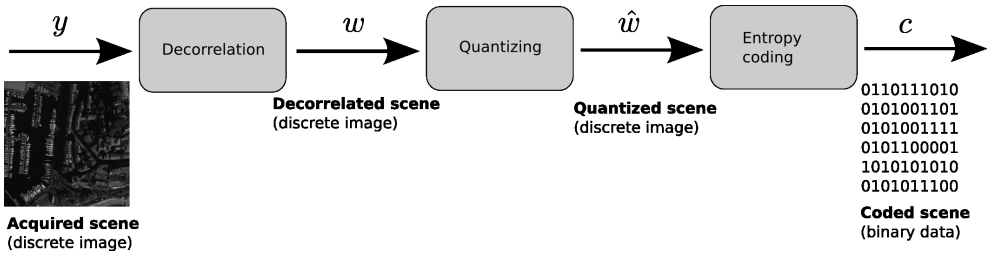


Figure 2.1: Stages of lossy image coding.

are named sparse transforms and provided autocorrelation matrices which tend to be diagonal. The optimal transform for the data decorrelation is the Karhunen-Loève transform[1] (KLT) as it provides a strict diagonal autocorrelation matrix. Its implementation is however difficult as the signal dependency of this transform makes it time-consuming to compute [Andrews 1971]. Until very recently, as on the SPOT 5 satellite, the discrete cosine transform (DCT), which is a signal-independent approximation of the KLT transform, was used [Wallace 1992].

Figure 2.2: Reference image, Cannes harbour (12 bits, 30 cm resolution, 1024×1024 pixels).

However, image quality evaluation of the DCT-based compression technique

[1] For signals which can be expressed as first-order Markov processes.

showed an acceptable compression rate of approximately 3 : 1 on SPOT5 remote sensing images (8 bits, 5 m resolution) [Thiebaut 2011], i.e. the compressed image is 3 times lighter than the original one. Higher compression rates wipe out the details of the image and create blocking artifacts on uniform zones. Such phenomenon is illustrated on Fig. 2.3 which shows the reference image (displayed Fig. 2.2) encoded at a rate of 2.5 bits/pixel (compression rate of almost 5 : 1). These artifacts appear because the DCT-based coding technique works on the image at a local level, i.e. on small 8×8 blocks. In order to bypass this compression bound for new generation high resolution satellites, like the PLEIADES-HR satellite, a new approach based on global transforms, such as the wavelet transform, has been adopted.

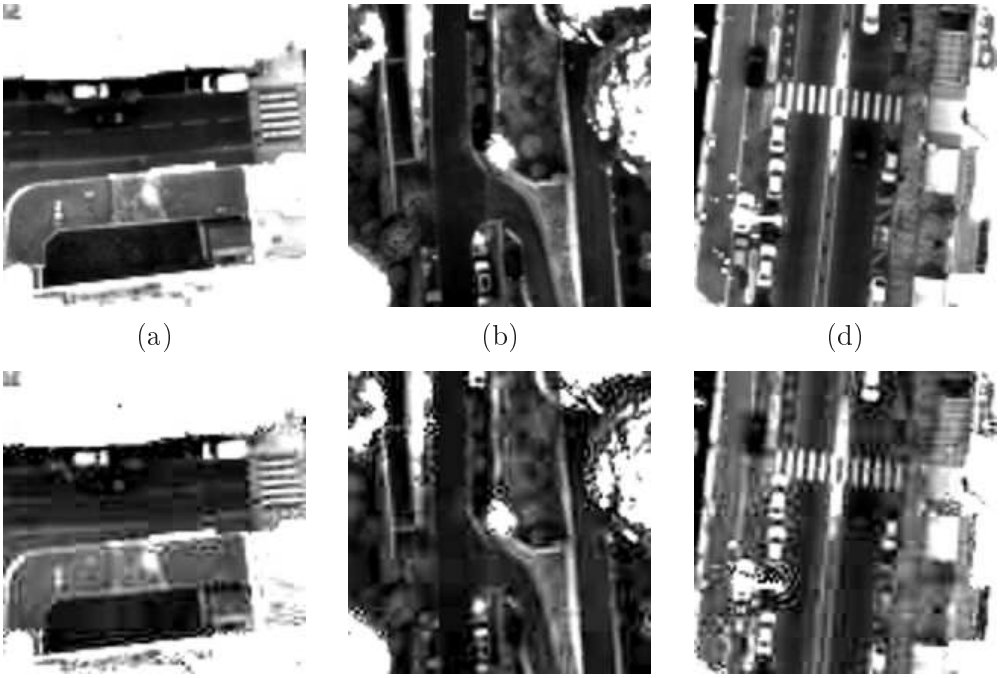


Figure 2.3: Visual comparison of the DCT-based compression technique. Displayed images have a size of 200×200 pixels. The first line shows zooms of different zones of the reference image. The second line represents the same zones but for the DCT-based decoded version of the reference image ($PSNR = 46.75\ dB$). The target rate is 2.5 bits/pixel (the dynamic range of the reference image is encoded on 12 bits). The image range has been extended to point out the image reconstruction artifacts.

Unlike the Fourier transform which is localized in frequency domain but not in spatial domain and the usual representation which is localized in spatial domain but not in frequency domain, the wavelet transform appears to be (more or less) localized both in space and in frequency. The multiresolution analysis algorithm proposed in [Mallat 1989] is recommended to process the wavelet transform of the image. This scheme is illustrated Fig. 2.4 for a one dimensional signal. It decomposes the image

in low and high frequencies by applying, in parallel, a low-pass filter h and a high-pass filter g both followed by subsampling operators. Two sets of coefficients are then obtained: The approximation coefficients which correspond to the low frequencies of the signal and which can be interpreted as a zoomed out version of the original signal and the details coefficients which correspond to the high frequencies of the signal. This decomposition process is then iterated on the approximation coefficients L times, L being referred as the number of levels decomposition.

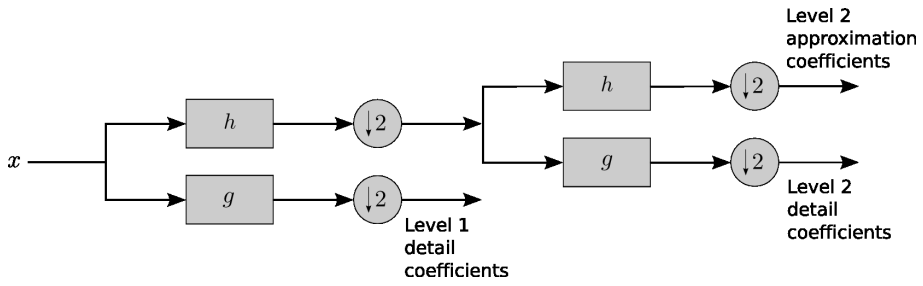


Figure 2.4: Filter banks for of a one level multiresolution analysis algorithm.

A wavelet transform can be extended to multidimensional signals using separable wavelets. Images can then be decomposed using the scheme described Fig. 2.4 iteratively on the rows and the columns of the image. The interested reader may refer to [Mallat 2008] for more details.

Different families of wavelets can be used for the decomposition of the image. The Cohen-Daubechies-Feauveau (CDF) 9/7 wavelet is often used in the image coding community as it owns interesting properties for image compression such as symmetric filters and enough number of vanishing moments which create short length filters while giving efficient sparse representations for most smooth images [Cohen 1992]. The definition of the corresponding filters h and g is given in Table 2.1 for the analysis of the image. Note that the CDF 9/7 wavelet transform is actually the wavelet transform recommended in the recent JPEG-2000 standard and is also the transform used by the PLEAIDES-HR satellite for image coding [Lier 2008].

k	Low-pass filter h_k	High-pass filter g_k
0	0.852698679009	−0.788485616406
±1	0.377402855613	0.418092273222
±2	−0.110624404418	0.040689417609
±3	−0.023849465020	−0.064538882629
±4	0.037828455507	

Table 2.1: Analysis filters for the 9/7 Cohen-Daubechies-Feauveau wavelet transform.

A wavelet transform is very sparse [Antonini 1992], meaning that it represents the image with a few number of non-null coefficients. This representation is very attractive for the encoders that follow the transform as they take benefit

of its sparsity to only encode the coefficients which bring information to the image and discard all the small wavelet coefficients. The statistical characteristics of a wavelet transform can also be taken into account to increase the coding performance [Shapiro 1993, Said 1996, Taubman 2000].

Once the image has been transformed, its coefficients need to be encoded to form the output bitstream. This encoding is usually done in two steps. The first step is the quantization of the coefficients which reduces the set of their values (usually reals) to a smaller set (usually integers). It also introduces a small correlation between the coefficients to improve the performances of the entropy coding that follows the quantization. This entropy encoding is then the second step of this process and converts the quantized coefficients into a binary stream. This conversion does not introduce any degradation and consequently is rarely displayed on coding schemes. The quantization is the part of the encoding process which introduces an irreversible degradation of the coefficients. This quantization can be explicitly performed as in the DCT-based compression system [Wallace 1992] or implicity, as the consequence of a bitstream truncature, for advanced encoders such as [Shapiro 1993, Said 1996, Taubman 2000]. We describe these encoders in the next lines. The encoder used on-board of current satellite imaging systems will be described in Chapter 3.

2.1.1.1 Embedded Zerotree Wavelet (EZW) encoder

The encoders proposed in [Shapiro 1993, Said 1996] are similar in the sense that they are both based on the hierarchical representation of a wavelet transform and exploit the self-similarity accross wavelet subbands (displayed Fig. 2.5). More precisely, the EZW encoder proposed in [Shapiro 1993] relies on the hypothesis that if a wavelet coefficient magnitude is below a given threshold T (it is said to be insignificant), then all the coefficients of the same orientation in the same spatial location at finer scales are likely to be insignificant too with respect to T. The EZW encoder then uses this hypothesis to create a significance map that only retain coefficients that bring information to the image.

This hierarchical notion allows to link the coefficients that belong to the same location and orientation together such that they can be represented by a *zerotree* structure. The objective of this structure is to locate the coefficients in the finer scales that are insignificant based on the magnitude of the coefficient currently scanned. The encoder can then predict the absence of significant coefficients at finer scales and stops the coding of the current tree. This technique is particularly efficient to quickly encode a wavelet transform as it contains many coefficients close to zero that do not bring much information to the image. This end-coding method is very similar to the end-of-block symbol used by the DCT-based compression system to stop the encoding of block when no more non-null coefficients are discovered. However in the EZW case, the encoder works on the whole image instead of small 8×8 blocks and therefore many more coefficients can be predicted to be insignificant using one symbol.

As mentioned earlier, the creation of the significance map depends on the value of

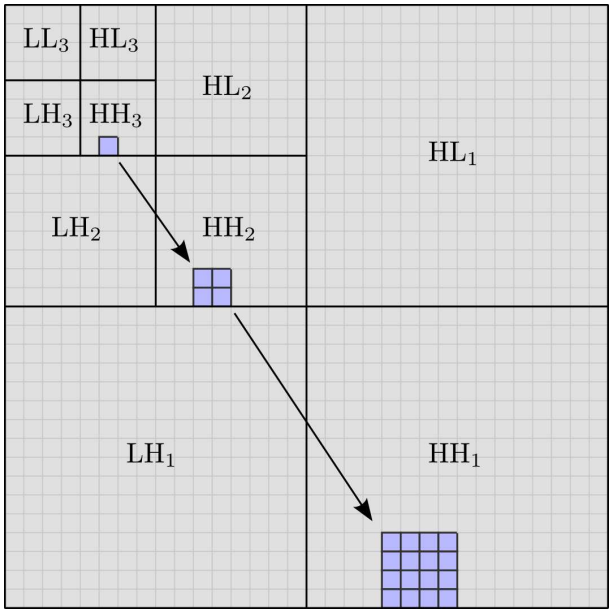


Figure 2.5: Coefficients dependencies through subbands on a 3-levels decomposition. The figure shows the same location at different scales.

the threshold T. In order to encode both large and small coefficients, this threshold needs to be decreased iteratively. This is the process of successive-approximation quantization (SAQ) [Shapiro 1993]. The SAQ creates a sequence of thresholds $T_i, i \in \{0, 1, \ldots, M\}$, where M is the number of iterations (usually set to the number of bits required to represent the maximum absolute value of the wavelet coefficients), and produces significances maps for each threshold T_i. Usually, the threshold at the iteration i is defined as the half of the previous threshold to match the binary representation of wavelet coefficients

$$T_i = \frac{T_{i-1}}{2}, \tag{2.1}$$

with T_0 is half of the first power of two greater than the maximum absolute value of the wavelet coefficients to encode.

During this iterative encoding procedure, two separate lists of wavelet coefficients are used to track the coefficients that have previously been marked as significant: The *dominant* and the *subordinate* lists. The dominant list contains the coordinates of the coefficients that have not been found to be significant yet while the surbordinate list contains the magnitudes of the coefficients that have been found to be significant.

The overall EZW algorithm is as follows. For each threshold, the dominant list is scanned and the significance map is produced. This map is then zerotree encoded using an algorithm described in [Shapiro 1993]. During this encoding, each coefficient marked as significant is removed from the dominant list and its magnitude is appended to the subordinate list. The coefficient is then set to zero in the data

to not disturb the computation of zerotrees of future iterations. Once the dominant list scan is completed, the magnitude of each coefficient of the subordinate list is refined. More precisely, a symbol is outputted to indicate if the true value of the encoded coefficient belongs to the upper or lower half of the current threshold. The encoding stops when the target rate has been reached.

The very good results obtained with this coder can be explained by the efficiency of the zerotree structure coupled with the SAQ technique, which in fact is almost equivalent to order the wavelet coefficient and to transmit first the large ones. This allows to decode the best possible image at any point in the binary stream: This is the process of progressive transmission. The Set Partitioning In Hierarchical Trees (SPIHT) encoder proposed in [Said 1996] is very similar to the EZW encoder as it also owns this feature of progressive transmission. However, the technique used by the SPIHT encoder to code the coefficients is radically different.

2.1.1.2 The Set Partitioning In Hierarchical Trees (SPIHT) encoder

The SPIHT algorithm also focusses on this aspect of progressive transmission but explicitly orders the wavelet coefficients and encodes first the large ones such that the mean square error (MSE) is minimized. Let y be the image to encode and $\hat{y}$ the decoded image, the MSE then writes

$$D = \frac{1}{N}\|y - \hat{y}\|^2, \tag{2.2}$$

where N is the number of pixels. Using an orthonormal wavelet transform, the definition of the MSE can be futher developped

$$D = \frac{1}{N}\|w - \hat{w}\|^2 = \frac{1}{N}\sum_{i=0}^{N-1}(w_i - \hat{w}_i)^2, \tag{2.3}$$

where w_i are the wavelet coefficients to encode and $\hat{w}_i$ the decoded coefficients. As mentioned in [Said 1996], it is clear that if the exact value of a coefficient w_i is transmitted, i.e. $\hat{w}_i = w_i$, then the MSE decreases by $\frac{w_i^2}{N}$. The SPIHT encoder is then based on the fact that the large wavelet coefficients need to be transmitted first so an image with the best quality (in the MSE sense) can be reconstructed at any time.

This encoder uses the binary representation of the coefficients and processes the data iteratively bit plane by bit plane through two passes: The sorting pass which orders the coefficients from the larger to the smaller and the refinement pass which outputs the bit value of current bit plane for each significant coefficient. The keypoint of this algorithm is that the coordinates of the sorted coefficients do not need to be transmitted as both encoder and decoder share the same execution path.

In detail, the strength of the sorting pass of the SPIHT encoder lies in the fact that it does not sort all coefficients but only selects the one that are significant with respect to a threshold T_n where n is the nth iteration (or sorting pass). To select these significant coefficients, the sorting pass divides all the pixels into partitionning

subsets $\mathcal{T}_m$ and evaluates the significance of each subset. If none coefficient of the subset $\mathcal{T}_m$ is significant, then the subset is considered as insignificant and is not processed any further. Otherwise, if at least one coefficient of the subset is significant, then the subset $\mathcal{T}_m$ is considered as significant and a specific rule is applied to divide the subset into new partition subsets $\mathcal{T}_{m,l}$ [Said 1996]. The significance test is then performed on these new subsets $\mathcal{T}_{m,l}$ and so on. This process is achieved iteratively until each subset is reduced to a single coefficient such that each coefficient has been found significant or not.

This significance map is then stored in three lists: The *list of insignificant sets* (LIS), *list of insignificant pixels* (LIP) and *list of significant pixels* (LSP). The LIP and LSP lists are used to respectively store the coordinates of insignificant and significant pixels. The LIS list is used to specify the type of subset associated to the coordinates of each coefficient.

The overall algorithm is as follows. It starts by initializing the number of iterations n to the number of bits required to represent the maximum value of the coefficients. For each entry of the LIP (which stores the coordinates of pixels which were evaluated as insignificant at the previous iteration), the significance is evaluated. The significant coefficients are moved to the LSP and their sign is outputted. The significance of the set of each entry of the LIS is then evaluated. If the set is found to be insignificant, it is added back to the LIS for the next iteration. Otherwise, it is further partionned. The resulting subsets are added back to the LIS and the single coefficient subsets are added either to the LIP or LSP depending on their significance. Each entry of the LSP is then processed by the refinement pass which outputs the nth most significant bits of the absolute value of the coefficients (the sign has already been outputted during the sorting pass). The value of n is then decremented by 1 to process the next bit plane.

As for the EZW encoder, the SPIHT encoder stops the encoding procedure once the bit budget has been exhausted. The quality can also be controlled by stopping the encoding procedure once the evaluation of (2.3) reaches the desired target value. Note that, contrary to the EZW algorithm, the SPIHT encoder directly produces the bitstream without using an entropy coding. As mentioned by [Said 1996], using an entropy coding does not bring much improvement and strongly increases the coding time. In the next part, we describe another well-known coding algorithm used in the JPEG-2000 standard.

2.1.1.3 Embedded Block Coding with Optimized Truncation (EBCOT) encoder

The JPEG-2000 standard is a recent recommendation for imaging coding and is also based on the wavelet transform described in Section 2.1.1. The JPEG-2000 entropy coder is based on the Embedded Block Coding with Optimized Truncation (EBCOT) contextual encoder proposed in [Taubman 2000]. This encoder is a block-based encoder organized in two layers named *Tiers*. The *Tier 1* divides each wavelet subband in small blocks and encodes each block using a contextual encoder.

The second layer, *Tier 2*, computes the optimal truncation points of the encoded bitstreams such that the global rate-distortion is minimized.

During the Tier 1, the encoder divides each wavelet subband into small 32×32 blocks and processes each block bit plane by bit plane. During this bit plane encoding procedure, the encoder scans each coefficient and processes through three different coding passes: The *Significance Propagation* pass, the *Magnitude Refinement* pass and the *Cleanup* pass. During each of these passes, four *primitives* are used: The Run-Length Coding (RLC) primitive, the Zero Coding (ZC) primitive, the Magnitude Refinement (MR) primitive and the Sign Coding (SC) primitive. These primitives are used to select the most appropriate context of the coefficient scanned depending on its neighbors. In detail, for each scanned coefficient, the eight adjacent neighbors are observed. Each neighboring configuration produces a specific context which is converted by the selected primitive to a particular output symbol. To limit the complexity of the coder, all the possible configurations have been reduced to eighteen contexts for all the primitives, one for the RL primitive, nine for the ZC primitive, five for the SC primitive and three for the MR primitive [Taubman 2000].

The coding of a bit plane is as follows. The *Significance Propagation* pass is used to locate the significant coefficients or the coefficients that have significant neighbors. Once these coefficients have been located, the RL and ZC primitives are invoked to identify the ones which become significant in the current bit plane. If so, the SC primitive is applied to encode their sign. During the *Magnitude Refinement* pass, the MR primitive is applied. This primitive is intended to refine the magnitude of the coefficients identified as significant by the *Significance Propagation* pass, by encoding the corresponding bits of the current bit plane. Finally, the *Cleanup* pass is used to encode the coefficients that have not been considered during the previous passes. The RL primitive is applied and the SC primitive is invoked if coefficients are found to be significant. Each outputted symbol is then encoded using an arithmetic coder.

Once each block has been encoded using the contextual encoder, the Tiers 2 computes the optimal truncation points of the encoded bitstream such that the truncation points lie on the rate-distortion convex hull. Let $D_i^{n_i}$ be the coding distortion of the block B_i whose bitstream has been truncated to the point n_i giving the coding rate $R_i^{n_i}$. As each block is encoded independently, the overall coding distortion D can be expressed as

$$D = \sum_{i=0}^{I-1} D_i^{n_i}, \tag{2.4}$$

where I is the number of blocks. Similarly, the overall coding rate R writes

$$R = \sum_{i=0}^{I-1} R_i^{n_i}. \tag{2.5}$$

The rate-distortion problem consists here in finding the optimal truncation points n_i^* which minimize the coding distortion D over the set $\mathcal{N}_i$ of all possible

truncation points, under the constraint that the coding rate R does not exceed the target rate R_c. It can be formalized as follows

$$n_i{}^* = \begin{array}{cc} \arg\min & \sum_{i=0}^{I-1} D_i^{n_i} \\ \text{subject to} & \sum_{i=0}^{I-1} R_i^{n_i} \leq R_c \\ & n_i \in \mathcal{N}_i \end{array} . \tag{2.6}$$

For some value of the Lagrange multiplier λ [Everett 1963], the problem (2.6) can be written in an unconstrained form [Taubman 2000]

$$n_i^{\lambda *} = \begin{array}{cc} \arg\min & \sum_{i=0}^{I-1} \left(D_i^{n_i^\lambda} + \lambda R_i^{n_i^\lambda} \right) \\ \text{subject to} & n_i^\lambda \in \mathcal{N}_i^\lambda \end{array} . \tag{2.7}$$

The rate-distortion optimization performed by the Tiers 2 consists thus in finding the value of λ such that the optimal truncation points $n_i^{\lambda *}$ in (2.7) satisfy $\sum_{i=0}^{I-1} R_i^{n_i^{\lambda *}} = R_c$. The optimization (2.7) can be performed numerically by finding, for a given λ, the minimal truncation point $j \in \{1, 2, 3, \dots\}$ which verifies for each block B_i

$$\frac{\Delta D_i^j}{\Delta R_i^j} = -\lambda, \tag{2.8}$$

where

$$\Delta D_i^j = D_i^{j-1} - D_i^j, \tag{2.9}$$

$$\Delta R_i^j = R_i^{j-1} - R_i^j. \tag{2.10}$$

Until now, the EBCOT encoder described here allows to reach the state-of-the-art image coding performances [Taubman 2000]. Its high computational cost make it difficult to use it on-board of a satellite. The encoders presented in [Shapiro 1993, Said 1996] are less expensive in term of computational resources and are frequently used as the basis of satellite embedded image coder (see Section 3.2.2).

2.1.2 Restoration techniques

In this part, we describe the techniques used for the restoration of the decoded image. Note that we only focus on the methods which decompose the restoration in a direct deconvolution followed by a threshold operation of some sparse representation. We do not include the methods based on a variational framework such as [Bect 2004] as they are time consuming to compute.

2.1.2.1 Wavelet thresholding estimators

Most of restoration techniques used in satellite imaging are based on the technique proposed in [Kalifa 2003b]. These methods consider that the observed image y is the result of the real scene x blurred by the point spread function (PSF) h of the optics and noised by an additive random noise n

$$y = h * x + n, \tag{2.11}$$

where $*$ denotes the convolution product. To simplify the notation, the sampling operation does not appear in the model (2.11) and we assume that all the variables are discrete.

The PSF h of the optics acts as a low-pass filter which attenuates the high frequencies of the image (edges and sharp textures) making it blurry. Retrieving the true image x from the observed one y is an ill-posed problem which requires prior information on the image x and on the noise n [O'Sullivan 1986]. As mentioned previously, one technique to address this problem is to formalize this estimation as a minimization problem using a variational approach. In detail, a variational approach consists in formulating the inverse problem as a minimization problem composed of a data fidelity term built from the noise model and a regularizing function suited to represent the image x [Chambolle 1997]. A general framework for the formulation of inverse problems using variational approaches has been proposed in [Bect 2004]. The resulting algorithms appear however to be quite time consuming and are thus inadapted to high resolution satellite imaging.

Here, we focus instead on methods similar to [Kalifa 2003b] which proposes to invert the problem (2.11) in two steps. The first step consists in dividing, in the Fourier domain, the observed image by the optical transfer function (OTF) to remove the attenuation of the filter h. This direct inversion tends however to amplify the noise, so the deconvolved image is usually decomposed in some sparse basis and its coefficients are then thresholded to reduce the energy of the amplified noise. These techniques belong to the class of thresholding estimators [Donoho 1994].

In the case of an image only degraded by an additive Gaussian noise, [Donoho 1994] showed that the maximum risk of these thresholding estimators is minimized if the vector basis of the decomposition concentrate the energy of the image over few coefficients and if the noise coefficients are nearly independent. It is well-known that wavelet basis own this property of sparsity as they are widely used for image compression [Antonini 1992]. As these transforms are orthogonal (or biorthogonal), the nearly independence between noise coefficients is achieved.

When the image is also degraded by blur, [Kalifa 2003b] showed that thesholding estimators based on wavelet basis may not be efficient as the deconvolved noise is colored. Let h^{-1} be the pseudo-inverse filter whose Fourier transform $\mathcal{F}(h^{-1})(u)$ is defined by

$$\mathcal{F}(h^{-1})(u) = \begin{cases} \frac{1}{\mathcal{F}(h)(u)}, & \text{if } \mathcal{F}(h)(u) \neq 0 \\ 0, & \text{otherwise} \end{cases}. \tag{2.12}$$

The deconvolved image $\tilde{x}$ is obtained by applying the pseudo-inverse filter h^{-1} to the observed image y

$$\tilde{x} = h^{-1} * y = w * x + z, \tag{2.13}$$

where z is the deconvolved noise and w is some regularizing function which cancels the frequency of the image where $\mathcal{F}(h)$ vanishes

$$\mathcal{F}(w)(u) = \begin{cases} 1, & \text{if } \mathcal{F}(h)(u) \neq 0 \\ 0, & \text{otherwise} \end{cases}. \tag{2.14}$$

The power spectrum S_z of the deconvolved noise z can be expressed as

$$S_z(u) = \begin{cases} \frac{S_n(u)}{|\mathcal{F}(h)(u)|^2}, & \text{if } \mathcal{F}(h)(u) \neq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{2.15}$$

From (2.15), we see that the power of the noise will be higher in the high frequencies where the magnitude of the Fourier transform of the filter h is low. A thresholding of some sparse decomposition is then required to reduce the intensity of the deconvolved noise. For deconvolution problems where the magnitude of the Fourier transform of the filter h decreases slowly, [Donoho 1995b] showed that wavelet basis still lead to efficient thresholding estimators for this class of deconvolution problems.

If the magnitude of the Fourier transform of the filter h vanishes, then theshlding in wavelet basis does not lead to satisfying results [Kalifa 2003b]. As the Fourier transform of h vanishes, the pseudo-inverse filter h^{-1} deals with important variations in the high frequency domain where the magnitude of the OTF goes near zero. Unfortunately, the high frequency subbands of wavelet basis do not have a sufficiently fine frequency resolution to concentrate the energy of the deconvolved noise in few coefficients. A wavelet packet decomposition [Coifman 1992] needs to be used to achieve an efficient estimation [Kalifa 2003b]. Hybrid Fourier-Wavelet approaches [Neelamani 2004] can also be used to deal with the frequencial representation of the colored noise.

A wavelet packet decomposition extends the discrete wavelet transform by iterating the decomposition both on the low frequency and the high frequency subbands. An exemple of such decomposition is illustrated Fig. 2.6 in comparison to a classical dyadic wavelet transform. We see that a wavelet packet transform leads to a representation with a finer frequency resolution in the high frequency subbands. For bounded variations signals, [Kalifa 2003a] showed that thresholding estimators based on wavelet packet decompositions are nearly minimax optimal for this class of deconvolution problems.

Thresholding estimators based on real wavelet packet transforms produce however artifacts on the reconstructed image. These artifacts come from the fact that real wavelet packet transforms suffer from a lack of shift invariance and a poor directionality. The lack of shift invariance can be worked around by applying the transform on shifted version of the deconvolved image. This however tends to significantly slow down the algorithm. The poor directionality comes from the fact that wavelet transforms are extended to the two-dimensional case using separable wavelets. This allows efficient decomposition algorithms which apply the wavelet transform independently on each dimension (rows and columns) of the image. Consequently, a two-dimensional wavelet transform only selects horizontal and vertical frequencies of the image but does not correctly represent the diagonal frequencies (oriented objects). This lack of directional selectivity creates aliasing artifacts which are particularly visible on the oriented objects (buildings, roads) of the reconstructed image.

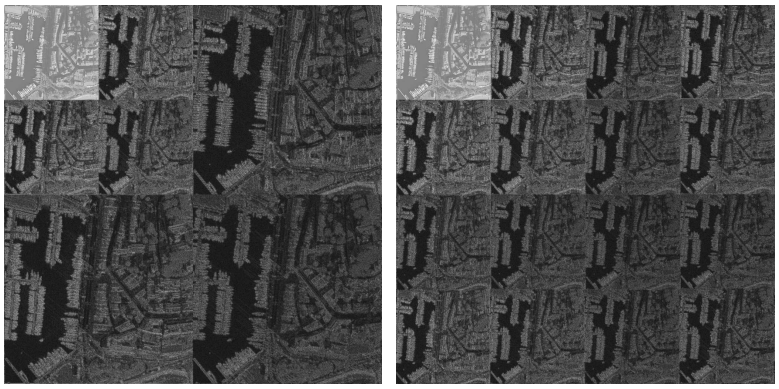


Figure 2.6: On the left, absolute value of a 2-levels wavelet decomposition of the reference image presented Fig. 2.2. On the right, absolute value of a 2-levels wavelet packets decomposition of the same image. Both transforms use orthogonal Daubechies DB6 filters set [Daubechies 1992].

Redundant wavelet transforms can be used to deal with the lack of shift invariance and the poor directionality, at the cost of more complex algorithms. An extension of the real wavelet packet transform to the complex case has been proposed in [Jalobeanu 2003]. This complex wavelet packet transform is based from the complex wavelet framework proposed in [Kingsbury 1998] which offers nearly shift invariance and a better directional selectivity with a limited redundancy. Advanced redundant wavelet transforms such as [Labate 2005] and [Candès 2005] can be used to capture specific features of the image (curves, oriented objects). Finally, note that all the referred methods can also take benefit of risk optimization techniques to estimate the optimal threshold parameters which minimize the MSE without the knowledge of the true image [Pesquet 2009], [Chesneau 2010]. A comparison of the state-of-the-art sparse transforms for image restoration will be presented in part II.

2.2 Joint optimization techniques

In this part, we briefly describe the methods which optimize one part of the imaging chain by taking into account the characteristics of the other components.

2.2.1 Optimal rate-allocation based models

To the best of our knowledge, the main contribution on joint optimization for image coding is the technique proposed in [Parisot 2002]. In this work, the authors proposed to compute a rate-allocation based on a wavelet subband model. The interesting point in the proposed method is that the global distortion can be weighted to take into account the post-processing steps. As explained in Section 2.1.2.1, the restoration done on-ground first performs a deconvolution to enhance the high frequencies of the image. It seems then interesting to weight the high frequency

subbands during the rate-allocation such that they are preserved from the quantizing [Parisot 2001]. More precisely, [Parisot 2002] proposed to write the coding distortion D as

$$D = \sum_{j=0}^{J-1} \Delta_j \pi_j D_j, \tag{2.16}$$

where J is the number of wavelet subbands, D_j is the coding distortion in the subband j and π_j are weighting coefficients which depend on the filters and the decimation factors used in the wavelet transform [Usevitch 1996]. Note that these weighting coefficients are only required if one considers biorthogonal wavelet transforms such as the CDF 9/7 wavelet transform [Cohen 1992]. They are equal to 1 for an orthogonal wavelet transform. The weighting coefficients Δ_j allow to favor one subband (i.e. one range of frequencies) during the rate-allocation problem. A low value of this weight will preserve the corresponding subband while a high value will penalize it.

Similarly, the coding rate R can be expressed as a function of each subband rate R_j

$$R = \sum_{j=0}^{J-1} a_j R_j, \tag{2.17}$$

where

$$a_j = \frac{N_j}{N}, \tag{2.18}$$

is the weight of the subband j in the whole image, that is the ratio between the size N_j of the subband j and the size N of the image.

The authors of [Parisot 2002] further proposed to modelize each wavelet subband using a centered generalized Gaussian distribution (GGD) law (the low frequency subband matches this model if a differential coding is first applied). Each subband is then parametrized by a standard deviation σ_j and a shape parameter α_j. Although several quantization models are considered in [Parisot 2002], each of them can be defined by the quantized step q_j and the size of a dead-zone z_j. A dead-zone is the quantizing interval which outputs a zero value. As shown in [Parisot 2002], using a dead-zone larger than the quantizing step gives better compression performances. The coding distortion D and the coding rate R can be expressed analytically as a function of the GGD and the quantization parameters [Parisot 2002]

$$D = \sum_{j=0}^{J-1} \Delta_j \pi_j \sigma_j^2 D_j \left(\alpha_j, \frac{z_j}{\sigma_j}, \frac{q_j}{\sigma_j} \right), \tag{2.19}$$

$$R = \sum_{j=0}^{J-1} a_j R_j \left(\alpha_j, \frac{z_j}{\sigma_j}, \frac{q_j}{\sigma_j} \right). \tag{2.20}$$

The rate-allocation problems consists here in finding the optimal quantizing parameters (quantizing step q_j^* and size of the deadzone z_j^*) which minimize the

coding distortion D under the constraint that the coding rate R does not exceed the target rate R_c

$$q_j^*, z_j^* = \begin{array}{ll} \arg\min & \sum_{j=0}^{J-1} \Delta_j \pi_j \sigma_j^2 D_j \left(\alpha_j, \frac{z_j}{\sigma_j}, \frac{q_j}{\sigma_j}\right) \\ \text{subject to} & \sum_{j=0}^{J-1} a_j R_j \left(\alpha_j, \frac{z_j}{\sigma_j}, \frac{q_j}{\sigma_j}\right) \leq R_c \\ & q_j, z_j \end{array} . \tag{2.21}$$

We immediately see that if $\Delta_j = 0$ then the corresponding subband j will not be included in the minimization of the distortion, leading to the minimal value of quantizing step ($q_j = 1$). High frequency subbands may then be preserved from excessive quantizing, which is preferable for the restoration that follows. The coding technique proposed in [Parisot 2002] is jointly optimized in this sense.

One can show that for some value of the Lagrange multiplier λ [Everett 1963], the rate-allocation problem (2.21) can be written in an equivalent unconstrained form

$$\tilde{q}_j^*, \tilde{z}_j^* = \begin{array}{ll} \arg\min & \sum_{j=0}^{J-1} \Delta_j \pi_j \sigma_j^2 D_j \left(\alpha_j, \tilde{z}_j, \tilde{q}_j\right) \\ & +\lambda \left(\sum_{j=0}^{J-1} a_j R_j \left(\alpha_j, \tilde{z}_j, \tilde{q}_j\right) - R_c\right) \\ \text{subject to} & \tilde{q}_j, \tilde{z}_j \end{array} . \tag{2.22}$$

where $\tilde{z}_j = \frac{z_j}{\sigma_j}$ and $\tilde{q}_j = \frac{q_j}{\sigma_j}$. Except under high coding rate assumption, the problem (2.22) cannot be solved in closed-form. The algorithm proposed in [Parisot 2002] to solve (2.22) is based on the resolution of the simultaneous equations obtained from the Karush-Kuhn-Tucker (KKT) conditions [Kuhn 1951] of problem (2.22). The KKT conditions are the necessary first order conditions for a solution of an optimization problem to be optimal. In clear, the KKT conditions state that the first derivatives of the function to minimize, taken at an optimal point, have to cancel. Note that these conditions are usually not sufficient and the analysis of the second derivatives is sometimes required to determine if the extremum found is a maximum, a minimum or a saddle point. Due to the complexity of problem (2.22), the authors of [Parisot 2002] explicitly assume that a mimimum exists and is unique. Only one point can then verify the KKT conditions of problem (2.22). These conditions write

$$\Delta_j \pi_j \sigma_j^2 \frac{\partial D_j}{\partial \tilde{z}_j}(\alpha_j, \tilde{z}_j^*, \tilde{q}_j^*) + \lambda^* a_j \frac{\partial R_j}{\partial \tilde{z}_j}(\alpha_j, \tilde{z}_j^*, \tilde{q}_j^*) = 0, \tag{2.23}$$

$$\Delta_j \pi_j \sigma_j^2 \frac{\partial D_j}{\partial \tilde{q}_j}(\alpha_j, \tilde{z}_j^*, \tilde{q}_j^*) + \lambda^* a_j \frac{\partial R_j}{\partial \tilde{q}_j}(\alpha_j, \tilde{z}_j^*, \tilde{q}_j^*) = 0, \tag{2.24}$$

$$\sum_{j=0}^{J-1} a_j R_j(\alpha_j, \tilde{z}_j^*, \tilde{q}_j^*) - R_c = 0. \tag{2.25}$$

The solution $\tilde{z}_j^*$ can be expressed as a function of the quantizing step $\tilde{q}_j^*$ and the shape parameter α_j, and therefore can be noted as [Parisot 2002]

$$\tilde{z}_j^* = g_{\alpha_j}(\tilde{q}_j^*). \tag{2.26}$$

Problem (2.22) is then reduced to find $\lambda^*, \tilde{q}_j^*$ which verify

$$h_{\alpha_j}(\tilde{q}_j^*) = -\frac{\lambda^* a_j}{\Delta_j \pi_j \sigma_j^2}, \tag{2.27}$$

$$\sum_{j=0}^{J-1} a_j R_j(\alpha_j, g_{\alpha_j}(\tilde{q}_j^*), \tilde{q}_j^*) = R_c, \tag{2.28}$$

where

$$h_{\alpha_j}(\tilde{q}_j) = \frac{\frac{\partial D_j}{\partial \tilde{q}_j}(\alpha_j, g_{\alpha_j}(\tilde{q}_j), \tilde{q}_j)}{\frac{\partial R_j}{\partial \tilde{q}_j}(\alpha_j, g_{\alpha_j}(\tilde{q}_j), \tilde{q}_j)}. \tag{2.29}$$

The monotonicity of functions h_{α_j} and R_j allows to solve numerically (2.27) and (2.28) using root-finding algorithms such as binary search procedures. From (2.29), we see that the function h_{α_j} only depends on the shape parameter α_j and $\tilde{q}_j$, the ratio between the quantizing step q_j and the standard deviation σ_j of the current subband. Without knowing explicitly the values of q_j and σ_j, one can numerically compute h_{α_j} for a given α_j and several values of $\tilde{q}_j$. Eq. (2.27) can then be solved using the generated lookup table (LUT) and a binary search procedure [Parisot 2002]. The same technique is applied to (2.28) to find λ^*. Solutions q_j^*, z_j^* and then deduced from $\tilde{q}_j^*$ and $\tilde{z}_j^*$ given σ_j.

In terms of coding performances, the technique proposed in [Parisot 2002] equals (and sometimes outperforms) JPEG-2000, which is the state-of-the-art of coding algorithm. The complexity of the algorithm [Parisot 2002] is however 5 times lower that JPEG-2000. These features make the method proposed in [Parisot 2002] to be very suitable for future high-resolution satellite compression scheme [Parisot 2000b].

2.2.2 Optimal joint decoding/deblurring

As mentioned in the Section 2.1.2.1, the restoration performed on-ground after decoding usually does not take into account the quantizing noise and considers the image formation model (2.11). But the coding step cannot be neglected at low coding rates and introduces a quantizing error. The method presented in [Tramini 1999] focusses on this aspect and proposes a restoration method which considers all the degradation of the imaging chain.

Let W be the wavelet transform used in the coding step, Q the quantizing operator and S the set of coordinates of the N pixels of the image. The quantized image $\hat{w}$ in the transformed domain writes

$$\hat{w} = Q\left(W\left(Hx + n\right)\right), \tag{2.30}$$

where x is the real scene, H is the matrix notation of a filtering process h (which stands as the PSF of the optics of the satellite) and n is the instrumental noise. In [Tramini 1999], the noise n is assumed to be centered, bounded, non-stationary and following a uniform distribution. But other considerations can be made to adapt the method to the considered chain. The variance σ_i^2 at the pixel i of the

instrumental noise relies on the value of the observed pixel $(Hx)_i$ and can be written as [Tramini 1999]

$$\sigma_i^2 = \alpha + \beta(Hx)_i + \gamma(Hx)_i^2, \quad \forall i \in S \tag{2.31}$$

where α, β and γ are three constants which depend on the acquisition parameters. The noise n is assumed to be uniformly distributed; its probability density function p_n can be expressed as a function of the variance σ

$$p_n(n) = \begin{cases} \frac{1}{2\sigma\sqrt{3}}, & \text{if } -\sigma\sqrt{3} \leq n < \sigma\sqrt{3} \\ 0 \end{cases} . \tag{2.32}$$

Each pixel noise $(n)_i$ is then bounded by

$$-\sigma_i\sqrt{3} \leq (n)_i < \sigma_i\sqrt{3}, \quad \forall i \in S. \tag{2.33}$$

Let b be the quantizing noise in the transformed domain, (2.30) can be written as

$$\hat{w} = W\left(Hx + n\right) + b. \tag{2.34}$$

Under the consideration that a subband uniform scalar quantizer is used, each pixel of a quantizing noise subband b_j is bounded by the quantized step q_j applied to the subband j

$$-\frac{q_j}{2} \leq (b_j)_i < \frac{q_j}{2}, \quad \forall i \in S_j. \tag{2.35}$$

where S_j is the set of coordinates of the N_j coefficients of the subband j. Eq. (2.34) can be further reduced to

$$\hat{w} = WHx + \varepsilon, \tag{2.36}$$

where $\varepsilon = Wn + b$. As mentioned in [Tramini 1999], the difficulty here is to bound the wavelet transform of a non-stationary noise. Under some stationary approximation of the instrumental noise in the transformed domain, the authors of [Tramini 1999] proposed to compute numerically the bound ω_j for each wavelet subband j of the instrumental noise. Each pixel of a subband ε_j of the global error ε then verifies

$$-\left(\frac{q_j}{2} + \omega_j\right) \leq (\varepsilon_j)_i < \left(\frac{q_j}{2} + \omega_j\right), \quad \forall i \in S_j. \tag{2.37}$$

From equation (2.37), one defines for each subband j the interval [Tramini 1998]

$$I_j = \left\{x \in \mathbb{R}^{N_j}, -\left(\frac{q_j}{2} + \omega_j\right) \leq (x)_i < \left(\frac{q_j}{2} + \omega_j\right), \forall i \in S_j\right\} \tag{2.38}$$

such that $\varepsilon_j \in I_j$. From (2.36), it is clear that

$$(\hat{w} - WHx) \in I, \tag{2.39}$$

where

$$I = \left\{x \in \mathbb{R}^N, x_j \in I_j, \forall j \in \{0, 1, \ldots, J-1\}\right\}, \tag{2.40}$$

where J is the number of wavelet subbands. The restoration method proposed in [Tramini 1999] is based on a variational approach and consists in minimizing the sum of two convex functions under the constraint that the global error belongs to I. This writes

$$x^* = \begin{array}{ll} \arg\min & f_1(x) + f_2(x) \\ \text{subject to} & (\hat{w} - WHx) \in I, \\ & x \in \mathbb{R}^N \end{array}, \tag{2.41}$$

where f_1 is the data fidelity term and f_2 is the regularizing term. The data fidelity term usually depends on the statistics of the noise. Here, the authors of [Tramini 1999] proposed to write the data fidelity term as

$$f_1(x) = \sum_{j=0}^{J-1} \sum_{i \in S_j} \frac{1}{2\sigma_j^2} \pi_j \left(WHx - \hat{w}\right)_i^2, \tag{2.42}$$

where σ_j^2 is the variance of the instrumental noise, approximated as stationary, in the subband j, and π_j are weightings coefficients required for biorthogonal wavelet transforms [Usevitch 1996]. The purpose of the regularizing term f_2 is to avoid the explosion of the noise during the deconvolution. It is built following some assumptions on the image. Here, the image is supposed to be a piecewise smooth function; the norm of its gradient is then assumed to be low [Rudin 1992]. The regularizing term proposed in [Tramini 1999] writes

$$f(x) = \sum_{i \in S} (G\lambda)_i \Psi \left(\left| \left(\nabla x \right)_i \right| \right), \tag{2.43}$$

where Ψ is an edge-preserving regularization function [Charbonnier 1997] and ∇ is the gradient operator. The regularizing term in (2.43) is controlled by the parameter λ which weights the regularization compared to the fidelity to the data. Usually, this parameter is a scalar such that the regularization is the same all over the image. The authors of [Tramini 1999] proposed to use a regularizing map (built by classification) such that the sensitive zones are not too smoothed. As this regularizing map is not differentiable, it is then smoothed using a convolution with a Gaussian kernel G. The minimization of the problem (2.41) is obtained from the numerical resolution, using the search method proposed in [Tramini 1998] and derived from [Uzawa 1958], of Euler-Lagrange equations associated to (2.41). As shown by the results of the method [Tramini 1999], taking into account the coding noise in the restoration allows to slightly improve the quality of the reconstructed image. The drawback of the method is that the prior used for the regularizing term tends to create flat homogeneous regions which are not appreciated from image analysis experts as they cannot be interpreted physically [Dherete 2003].

Chapter 3

Current CNES earth observing imaging chain systems

In this part, we describe the composition of a satellite imaging chain. A simplified representation of this chain is displayed figure 1.1. The role of each component of the imaging chain has already been described in Chapter 1. We focus in this chapter to the technical features of each of these components. The data presented in the thesis are provided by the CNES and are simulations of the post-PLEIADES new generation high-resolution satellites. We then focus only on the imaging chain system used by the CNES but the methods we propose are more general and can be easily extended to the characteristics of other chains.

3.1 Characteristics and instrument model

3.1.1 Transfert function of the optics

The optics of a satellite is built from a complex combination of mirors. The light emitted from the scene is reflected by these successive mirors and is then focalized on the detector. Several design of optics exist, such as the Korsch telescope wich is a three mirors telescope. The characteristics of the telescope depends on specifications such as the magnitude of the optical transfer function or the target sampling rate. For example, the PLEIADES-HR satellite uses a Korsch telescope [Lier 2008] with a 65 cm pupil of 12.9 m focal length. It allows to capture panchromatic images with a resolution of 70 cm and multispectral images with a resolution of 2.80 m. For the post-PLEIADES new generation satellites, a target resolution of 30 cm is planned.

The acquired signal is processed as follows. It is first sampled and transmitted to the eletronic parts to be shaped such that it is not too noisy. The signal is later amplified to fit all the available range and to limit the effect of the quantization during the analog-to-digital conversion. The analog-to-digital converter is the last part of the acquisition process. It quantizes the amplified signal on 12 bits, giving a digital image whose pixels vary from 0 to 4095.

This acquisition process affects the quality of the true image by adding blur and instrumental noise. The blur is mainly caused by the natural environment and the imperfection of the acquisition components. The atmosphere, the optics and the sensor all own a transfer function which attenuate the high frequencies of the image (edges, sharp textures) making it blurry. Let h_a, h_o and h_d respectively be the transfer functions of the atmosphere, the optics and the sensor. We assume that

all these operators are linear and translation invariant. The global point spread function h is then the convolution product of all the intermediate transfer functions

$$h = h_a * h_o * h_d. \tag{3.1}$$

Note that the Fourier transform of the global PSF, namely the optical transfer function, does not cancel at the Nyquist frequency and thus adds aliasing on the image. This aliasing phenomenon remains however limited as the magnitude of the optical transfer function (the MTF) at the Nyquist frequency is usually low. For example, the MTF is equal to 0.1 at the Nyquist frequency on the PLEIADES-HR satellite. This characteristic is one of the major point of the specifications of satellite optics.

3.1.2 Instrument noise model

The instrumental noise is also the composition of several noise sources such as a photon noise, an electronic noise and a quantizing noise due to analog-to-digital conversion. It is assumed to be centered and Gaussian with a variance σ_i^2 which depends on the observed pixel. Let $\sigma_{p_i}^2, \sigma_{e_i}^2, \sigma_{q_i}^2$ be respectively the variances of the photon noise, the eletronic noise and the quantizing noise at the pixel i. The variance of the global noise σ_i^2 at this pixel is expressed as the sum of the variances of the different noises

$$\sigma_i^2 = \sigma_{p_i}^2 + \sigma_{e_i}^2 + \sigma_{q_i}^2. \tag{3.2}$$

By taking into account the mathematical expression of each variance, one can approximate the variance σ_i^2 of the global noise at the pixel i as a linear function of the observed luminance $h * x$ sampled at the same pixel i [Lier 2008]

$$\sigma_i^2 = \alpha^2 + \beta(h * x)_i, \tag{3.3}$$

where α and β are two given constants (i.e. not pixel dependent). These two constants rely on the target signal-to-noise ratio (SNR) (which is function of the luminance) and directly derive from the parameters of the electronic chain such as the amplification factor or the quantizing step of the analog-to-digital converter. Two target luminances are usually used to compute the value of α and β: The mean luminance of the image, namely $L2$, which is defined as 97 W.m^{-2}.sr^{-1}.μm^{-1} and the luminance $L1$ defined as 14 W.m^{-2}.sr^{-1}.μm^{-1}. These luminances can be converted in pixel values by multiplying them by the ratio between the pixel maximum value (4095) and the maximum luminance value (370 W.m^{-2}.sr^{-1}.μm^{-1}). In pixels values, these luminances are then defined as $L1 = 154.94$ and $L2 = 1073.54$. Given the target signal-to-noise ratios associated to $L1$ and $L2$, one deduces the standard deviation of the global noise at the two target luminances

$$\sigma_{L1} = \frac{L1}{SNR(L1)}, \tag{3.4}$$

$$\sigma_{L2} = \frac{L2}{SNR(L2)}. \tag{3.5}$$

From equation (3.3), we also have

$$\sigma_{L1}^2 = \alpha^2 + \beta L1, \tag{3.6}$$
$$\sigma_{L2}^2 = \alpha^2 + \beta L2. \tag{3.7}$$

Using (3.4) and (3.6), one can compute the constants α and β. Table 3.1 shows the values of these constants for several operating points (OP) simulated by the CNES on the reference image presented Fig. 3.1.

	OTF	Resolution	Coding rate	SNR (L1-L2)	α	β
OP 61	0.1	30 cm	4.0 bpp	30-100	3.2866	0.097780
OP 62	0.1	30 cm	2.5 bpp	30-100	3.2866	0.097780
OP 63	0.1	30 cm	4.0 bpp	30-150	4.6220	0.028128
OP 64	0.1	30 cm	2.5 bpp	30-150	4.6220	0.028128
OP 65	0.1	30 cm	4.0 bpp	50-150	1.5286	0.045790
OP 66	0.1	30 cm	2.5 bpp	50-150	1.5286	0.045790

Table 3.1: Parameters of the acquisition chain for several simulated operating points (OP). The column OTF displays the value of the OTF at Nyquist frequency. The column coding rate indicates the number of bits per pixel (bpp) achieved at the output of the compression algorithm.

Finally, we can modelize the discrete acquired image y (considered as a vector of length N, where N is the number of pixels) at the output of the acquisition chain as the convolution product of the real analog image x and the global PSF h (3.1), sampled on a grid Δ, and noised by the discrete instrumental noise n. This writes

$$y = (h * x)_\Delta + n. \tag{3.8}$$

We assume the grid Δ to be the usual square sampling grid. The variable h now refers to the discretization of the analog PSF on the grid Δ and x represents the convolution of the analog image with a target PSF (see Section 3.3.2), sampled on Δ. Note that this image x is the closest discrete approximation of the true analog image that we can obtained. Model (3.8) rewrites

$$y = h * x + n. \tag{3.9}$$

The instrumental noise n is assumed to follow a normal zero-mean distribution whose variance σ_i^2 at the pixel i depends on the observed pixel and is given by the model (3.3).

3.2 On-board processing: Image coding

Once the image has been acquired, it needs to be compressed for an efficient storage and transmission. The compression system embbeded on-board of PLEIADES-HR satellite processes the image in three steps, similarly to the coding scheme depicted

Figure 3.1: Reference image, Cannes harbour (12 bits panchromatic image, 30 cm resolution, 1024×1024 pixels).

Fig. 2.1. A wavelet transform is first applied to the image to reduce its correlation. A bit plane encoder is then used to encode the transformed data. The encoded coefficients are then converted by an entropic encoder to form the binary stream. We detail each step in the following.

3.2.1 Wavelet transform

Section 2.1.1 showed that the state-of-the-art image coding algorithms use wavelet transforms to decorrelate the data. Based on this observation, the Consultative Committee for Space Data Systems (CCSDS), which produces system standards for spaceflight, proposed a new image coding recommandation based on a wavelet transform [CCSDS 2005]. For example, the coding scheme of the PLEIADES-HR satellite highly relies on the latter. For implementation issues, the wavelet transform is however performed "on the fly" [Parisot 2000b] on-board of this satellite. The recommandation [CCSDS 2005] is very close to the SPIHT encoder and uses a three levels Cohen-Daubechies-Feauveau (CDF) 9/7 wavelet transform [Cohen 1992] followed by a bit plane encoder (BPE).

The purpose of the bit plane encoder proposed by the CCSDS consists in encoding the binary representation of the wavelet coefficients through a successive process of the bit planes. This encoder is described in the next part.

3.2.2 CCSDS Bit plane encoder (BPE)

The encoder proposed by the CCSDS is similar to the encoders EZW and SPIHT. It exploits the hierarchical representation of the wavelet transform to proceed first with the coefficients that bring information to the image. It is however a simplified version of these encoders to match the limited computing resources available on-board.

3.2.2.1 Structure of the BPE

Once the wavelet transform is completed, the coefficients are first rounded to the nearest integers and are then divided in *blocks* of 64 coefficients each (the composition of a block is detailed in Section 3.2.2.3). Fig. 3.5 displays this notion of block arrangement. We see that, in order to form a block, the encoder selects the same geographical zone for each frequency bands of each decomposition level. The purpose of this block arrangement is then to represent the same spatial zone for different frequency bands. This allows to control the encoding of a zone depending on its frequency content. A homogeneous zone may require less high frequencies than the zone covering the edges of a building, for example. A block arrangement is then efficient in this sense.

A block is composed of a single low frequency coefficient and 63 high frequency coefficients taken accross the high frequency subbands. To increase the coding performances of the encoder, S blocks are gathered into a *segment*. The image is then processed segment by segment. Usually the number of blocks S is chosen such that

a segment represents a thin horizontal strip of the image. In that case, a *strip* compression is performed [Yeh 2005]. This type of compression is efficient for memory limited implementations.

The overall procedure of a segment encoding is given in Table 3.2. For each segment, the encoder starts by producing a segment header. This header includes important information on the coding parameters and is therefore required for the decoding. This step is not detailed here but can be found in [Yeh 2005]. The second step of the procedure consists in encoding the low frequency coefficients. Due to the major role that play these coefficients in the wavelet recomposition algorithm, they should remain the most unchanged as possible. A specific encoding rule is consequently applied on these coefficients. Section 3.2.2.2 is dedicated to this aspect.

The last step of the segment coding procedure consists in encoding the bit planes of the high frequency coefficients from the most significant bit plane (MSB) to the least significant bit plane (LSB). A bit plane b is a binary image created from the b^{th} bit of the two's-complement binary representation of each low frequency coefficient and the b^{th} bit of the binary representation of each high frequency coefficient. To illustrate this notion of bit plane, let us consider the block displayed Fig. 3.2.

63	-34	49	10	7	13	-12	7
-31	23	14	-13	3	4	6	-1
15	14	3	-12	5	7	3	9
-9	-7	-14	8	4	-2	3	2
-5	9	-1	47	4	6	-2	2
3	0	-3	2	3	-2	0	4
2	-3	6	-4	3	6	3	6
5	11	5	6	0	3	-4	4

Figure 3.2: Illustration of a block.

The coefficient in grey is the low frequency coefficient and will be ignored for this example. We see that the highest coefficient among the high frequency coefficients is equal to 49. There are then 6 bit planes to encode as the highest coefficient is greater than $2^5 = 32$ but lower than $2^6 = 64$ (for that follows, the least significant bit will be referred to the zeroth bit). The first bit plane is $b = 5$ (the MSB). This bit plane is formed by the value of the fifth bit of each coefficient. On this example, only three coefficients have a fifth bit: -34, 49 and 37. The fifth bit plane is then the binary image composed of the value of the fifth bit of these coefficients (respectively -1, $+1$ and $+1$, the sign is also taken into acount). This give the binary image displayed Fig. 3.3.

Bit plane encoders are particularly efficient to encode signals when resources are

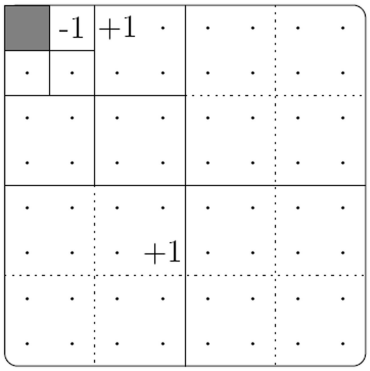


Figure 3.3: Illustration of the fifth bit plane of the block presented Fig. 3.2.

limited [CCSDS 2005]. Section 3.2.2.3 described the technique used to achieve this bit plane encoding.

Produce segment header
Encode low frequency coefficients
Encode bit plane $b = b_{max} - 1$ (MSB)
Encode bit plane $b = b_{max} - 2$
...
Encode bit plane $b = 0$ (LSB)

Table 3.2: Segment encoding procedure, b_{max} is the number of bit planes required to encode the magnitude of high frequency coefficients.

3.2.2.2 Coding of the low frequency coefficients

Preserving the low frequency coefficients from excessive quantizing is vital to reconstruct an image with a satisfying visual quality. As they initialize the wavelet recomposition algorithm, an error on the low frequency coefficients has an important impact on the fidelity of the decoded image. But their magnitude is very high (higher than the magnitude of the high frequency ones). A lossless encoding of these coefficients may then consume a lot of the bit budget, especially when the target rate is low. To allow this case, the encoder considers that the least significant bit planes of the low frequency coefficients can be slightly deteriorated without impacting the quality of the decoded image.

The encoder then processes the low frequency coefficients through a lossless encoding of their most significant bit planes using an explicit quantization followed by a differential coding scheme [CCSDS 2005]. The quantization step is chosen as a power of two such that the quantization of the coefficients is equivalent in shifting their bit planes. The remaining bits are then represented on b_{max} bit planes and are included in the bit plane encoding procedure described in Section 3.2.2.3.

3.2.2.3 Bit planes encoding

The bit planes encoding is the last step of the segment coding procedure. It processes the coefficients bit plane by bit plane following the five stages procedure depicted on Fig. 3.4. Each bit plane is processed separately. The blocks inside a bit plane are also treated independently one by one. This segment coding procedure is displayed Fig. 3.4.

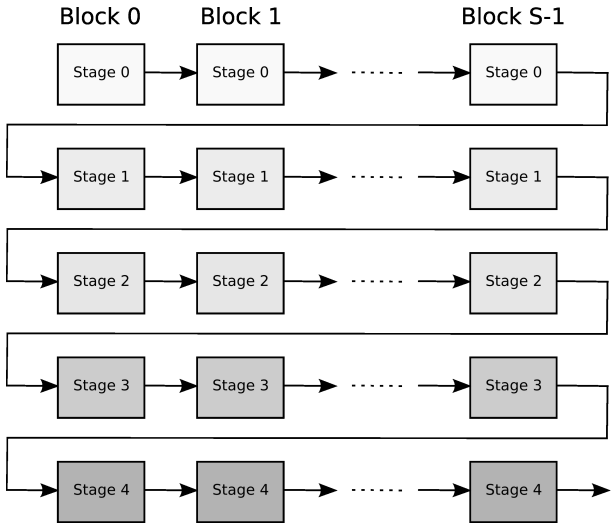


Figure 3.4: Encoding procedure of a bit plane.

The stage 0 simply consists in appending the b^{th} bit of the remaining bits of the low frequency coefficients to the output bitstream. The stages 1 to 4 are dedicated to the encoding of the high frequency coefficients.

The technique used to encode these coefficients is very similar to the technique proposed in [Shapiro 1993] which is based on the hierarchical representation of the wavelet transform to encode trees of non significant coefficients with respect to a threshold T. Here the encoder relies on the binary representation of the coefficients, processed bit plane by bit plane. We directly deduce that the threshold T is implicity the decimal value associated to the current bit plane b and is equal to 2^b. To evaluate if the coefficients are significant, the BPE simply tests their magnitude. It produces a codeword $t_b(w_i)$ named *type* which indicates if the scanned coefficient w_i has already been found significant in the previous bit plane (type 2), becomes significant in the current bit plane (type 1) or is not significant (type 0). The rule is as follows

$$t_b(w_i) = \begin{cases} 0 & \text{if } |w_i| < 2^b \\ 1 & \text{if } 2^b \leq |w_i| < 2^{b+1} \\ 2 & \text{if } 2^{b+1} \leq |w_i| \end{cases} . \tag{3.10}$$

At the bit plane b, only the coefficients which become significant (type 1) are encoded in stage 1-3. The coefficients whose type is 0 are not significant yet and are

passed over. The coefficients evaluated as type 2 have already been found significant in the previous bit planes and have therefore been already encoded in stages 1-3. The encoder just needs to refine their magnitude by appending their b^{th} bit to the output bitstream. This is the stage 4 of the process. To reach high compression rates, the BPE uses the same technique as [Shapiro 1993] and sets up a tree structure to efficiently encode trees of non significant wavelet coefficients. This tree is built using the block arrangement displayed Fig. 3.5.

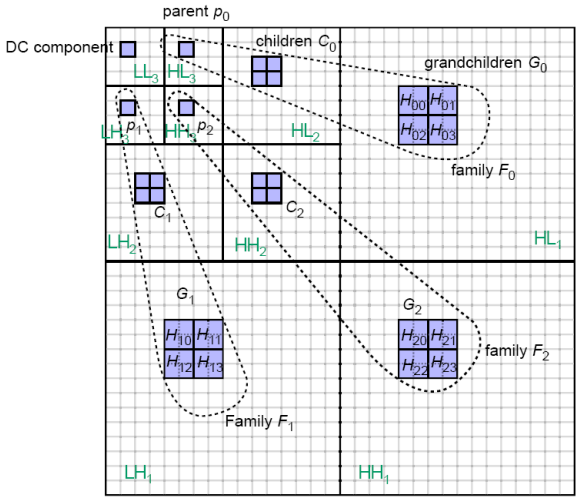


Figure 3.5: Wavelet block arrangement. This illustration is the property of the CCSDS.

A block is composed of one low frequency coefficient and 63 high frequency coefficients. To ensure some frequency selection, these 63 coefficients are partioned into three families F_0, F_1 and F_2. A family represents the same spatial information through the three (as the wavelet transform is performed on three levels of decomposition) different scales. Each family F_i is then made of

- One *parent* coefficient p_i.
- A group of four *children* coefficients C_i.
- A group of sixteen *grandchildren* coefficients G_i partitioned into four groups $H_{ij}, j \in \{0, 1, 2, 3\}$.

This family hierarchy is similar to the zerotree structure of the EZW encoder [Shapiro 1993] and is used to efficiently detect trees of non significant coefficients. These non significant trees can then be encoded using few bits, allowing to reach high compression rates. To encode these families, several lists are defined

- The list of parents $P = \{p_0, p_1, p_2\}$. For example, the list of parents corresponding to the block presented Fig. 3.2 is $P = \{-34, -31, 23\}$.

- The list of descendants D_i in a family i which includes the children and the grandchildren coefficients, $D_i = \{C_i, G_i\}$.
- The list of descendants B of a block which includes the descendants lists of all families, $B = \{D_0, D_1, D_2\}$.

For each bit plane b, the BPE encodes the coefficients which become significant (type 1) of the three families using a three stages procedure. Stage 1 scans the parents list P and evaluates the significance of each parent using the function (3.10). It then produces two *codewords* $types_b[P]$ and $signs_b[P]$. Let L be a list of coefficients

- $types_b[L]$ is the binary codeword consisting of the b^{th} magnitude bit of each coefficient w_i of L such that $t_b(w_i) \in \{0, 1\}$.
- $signs_b[L]$ is the binary codeword consisting of the sign bit of each coefficient w_i of L such that $t_b(w_i) = 1$. The sign of a coefficient is only coded once at the bit plane it becomes significant. The sign of a negative coefficient is represent by a 1 and the sign of a positive coefficient is represent by a 0.
- Given a list of types values $T = \{t_0, t_1, \ldots, t_l\}$, $tword[T]$ is the binary codeword consisting of the sequence of type values t_i that verify $t_i \in \{0, 1\}$.

On the parents list of the example block displayed Fig. 3.2, we have $t_5(-34) = 1$ (since 34 verifies $32 < 34 < 64$), $t_5(-31) = 0$ and $t_5(23) = 0$. Therefore only the coefficient equal to -34 is significant for the bit plane $b = 5$, so the BPE produces $types_b[P] = \{1, 0, 0\}$ and $signs_b[P] = 1$ (-34 is negative).

Once the BPE has scanned the parents list, it seeks some significant descendants. This is the stage 2. It first looks if there is any significant coefficient among the children and the grandchildren. It produces the $tran_B$ codeword

$$tran_B = \begin{cases} \varnothing & \text{if } tran_B = 1 \text{ at any previous bit plane } b \\ 1 & \text{if } \exists w_i \in B, t_b(w_i) = 1 \\ 0 & \text{otherwise} \end{cases} .$$

This transition codeword may be difficult to grasp and needs further explanations. The idea of the codeword $tran_B$ is to indicate if there exists at least one significant descendant. To do so, the BPE tests the significance of each coefficient that belongs to the descendants list B. If a coefficient w_i is found significant, the test function $t_b(w_i)$ will be equal to 1, 0 otherwise. The BPE takes then the maximum value over all significance tests to generate $tran_B$. If at least one descendant is significant, the BPE will then produces $tran_B = 1$. Note that this codeword is not generated if it has been previously produced equal to 1. It is indeed useless to generate this codeword for each bit plane if the BPE has already mentioned that significant descendants exist.

On the example illustrated Fig. 3.2, two descendants are significant for the bit plane $b = 5$ (49 and 47). The BPE produces then $tran_B = 1$.

Once at least one descendant has been found significant, one needs to locate in which family this descendant is. The BPE produces the codeword $tran_D$ to achieve this goal

$$tran_D = tword\Big[\{\max(t_b(D_i))\},\forall i \in \{0,1,2\} \text{ such that } \max(t_b(D_i)) \neq 1 \text{ in previous bit planes}\Big].$$

The behavior of this codeword is similar to $tran_B$: It indicates in which family i the descendants have been found. This codeword is not produced if $tran_B = 0$, meaning that there does not exist any significant descendants. The last step of the stage 2 is to produce the magnitude $types_b[C_i]$ and the sign $signs_b[C_i]$ codewords of the significant children. Note that the BPE only encodes the children of the families that have been marked as significant by the $tran_D$ codeword.

On the example Fig. 3.2, two descendants are significiant for the bit plane $b = 5$. These coefficients belongs to the descendant lists D_0 (for the coefficient 49) and D_1 (for the coefficient 47). We then have $tran_D = \{1, 1, 0\}$. As $tran_D$ has been generated, the BPE looks for some descendants in the corresponding children groups C_0 and C_1. The coefficient 49 is the zeroth bit of the children group C_1 while the coefficient 47 does not belong to the children group C_1 (but it belongs to one of the grandchildren groups which are processed in stage 3). Therefore the BPE produces $types_b[C_0] = \{1, 0, 0, 0\}$, $types_b[C_1] = \{0, 0, 0, 0\}$ and $signs_b[C_0] = 0$ (49 is postive). Codeword $signs_b[C_1]$ is empty because no coefficients have been found significant in the children group C_1.

The stage 3 is dedicated to the encoding of the grandchildren. Of course, this stage is omitted if the BPE produced $tran_B = 0$ at stage 2 implying that it is not necessary to look for significant grandchildren. Similarly to stage 2, stage 3 produces the codeword $tran_G$ to indicate in which family one may find significant coefficients

$$tran_G = tword\Big[\{\max(t_b(G_i))\},\forall i \in \{0,1,2\} \text{ such that } \max(t_b(D_i)) > 0 \text{ in current or previous bit planes}\Big].$$

As the grandchildren G_i of each family are further partitioned into four groups $H_{ij}, j \in \{0, 1, 2, 3\}$, the BPE needs to produce one more transition codeword to locate the significant coefficients

$$tran_H = tword\Big[\{\max(t_b(H_{ij}))\},\forall j \in \{0,1,2,3\}\Big] \forall i \in \{0,1,2\}.$$

The last step of the stage 3 is to produce the magnitude $types_b[H_{ij}]$ and the sign $signs_b[H_{ij}]$ codewords of the significant grandchildren. Again, note that the BPE only encodes the grandchildren of the families that have been marked as significant by the $tran_G$ and $tran_H$ codewords.

On the example Fig. 3.2, the BPE has already produced, during stage 2, $tran_D = \{1, 1, 0\}$ meaning that significant coefficients exist in families 0 and 1. the BPE

now looks if these significant coefficients belong the grandchildren groups of these families. At the bit plane $b = 5$, the coefficient 47 belongs to a granchildren group of the family 1. No coefficient are significant in the granchildren groups of the family 0. Therefore, the BPE produces $tran_G = \{0, 1\}$ and $tran_{H_1} = \{0, 1, 0, 0\}$ since the coefficient 47 is the bit 1 of the H_{11} group. We also have $types_b[H_{11}] = \{0, 1, 0, 0\}$ and $signs_b[H_{11}] = 0$ (47 is positive).

Table 3.3 summarizes the generated codewords. To form the final output bitstream, these codewords are encoded by a variable length entropy coder. As mentioned previously, the last stage (stage 4) of the coding procedure consists in including the b^{th} magnitude bit of each type 2 high frequency coefficient. If the target compression rate does not allow a lossless coding of the wavelet coefficients, the encoder truncates the output bitstream of each segment to reach the target rate. The coder also provides a quality control which consists of setting a maximum number of bit planes to encode. This option does not allow however to control the compression rate.

Stage 1 (parents)	$types_b[P]$, $signs_b[P]$
Stage 2 (children)	$tran_B$ $tran_D$ $types_b[C_i]$, $signs_b[C_i]$
Stage 3 (grandchildren)	$tran_G$ $tran_{H_i}$ $types_b[H_{ij}]$, $signs_b[H_{ij}]$

Table 3.3: Generated codewords for each coding stage.

3.3 On-ground processing: Image decoding and restoration

3.3.1 Image decoding and reconstruction

Once the bitstream has been transmitted, the decoder needs to reconstruct the image. The bitstream may have been truncated due to some coding rate constraint. To reconstruct the image, the decoder first completes the bitstream by adding zeros bits and then applies the inverse of the coding procedure described in Section 3.2.2.3. An inverse wavelet transform is then applied on the decoded coefficients to reconstruct the image.

The inverse transform scheme used to reconstruct the image is also based on the multiresolution analysis proposed in [Mallat 1989]. The obtained algorithm is illustrated on Fig. 3.6. This scheme is initialized with the low frequency coefficients of the decoded signal. These coefficients are upsampled and filtered by the low-pass filter $\tilde{h}$. The same process is applied to the details coefficient of the last decomposition level with the high-pass filter $\tilde{g}$. These filters are given in Table 3.4.

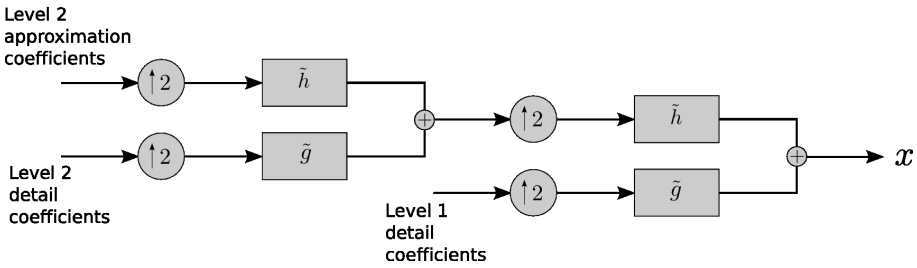


Figure 3.6: Filter banks for of an one level multiresolution synthesis algorithm.

The obtained two sets of coefficients are later added to reconstruct the signal. This reconstructed signal is then used as the intialization of the next level recomposition and so on. This process is iterated L times (L is the number of levels decomposition fixed to 3 in the case of the CCSDS recommandation) until all levels have been reconstructed.

k	Low-pass filter $\tilde{h}_k$	High-pass filter $\tilde{g}_k$
0	0.788485616406	−0.852698679009
±1	0.418092273222	0.377402855613
±2	−0.040689417609	0.110624404418
±3	−0.064538882629	−0.023849465020
±4		−0.037828455507

Table 3.4: Synthesis filters for the 9/7 Cohen-Daubechies-Feauveau wavelet transform.

Similarly to the decomposition scheme, the recomposition algorithm can be extended to two dimensional signals using the scheme described Fig. 3.6 iteratively on the rows and the columns of the image.

Once the image has been decoded and reconstructed, it needs to be restored. Indeed, at this point, the reconstructed image contains all the accumulated degradations of the imaging chain such as blur, instrumental and quantizing noise; the step of restoration is then crucial to produce an image which can be exploited.

3.3.2 Deconvolution and denoising

The restoration technique used by the CNES to improve the quality of the decoded image is based on the method proposed in [Kalifa 2003b] and described in Section 2.1.2.1. The restoration is then performed in two steps: The decoded image is first deconvolved to reduce the blur of the optics and is then denoised to limit the growth of the instrumental noise power due to the deconvolution. The acquisition model considered by the restoration method of the CNES is the same than the one used in [Kalifa 2003b] and writes

$$\hat{y} = h * x + n, \tag{3.11}$$

where $\hat{y}$ is the decoded image, x is the real scene, h is the PSF of the optics described in Section 3.1.1, and n is the instrumental noise whose model is given in 3.1.2. Note that the coding noise is not considered in this model. The deconvolution technique used by the CNES is slightly different from the one proposed in [Kalifa 2003b]. Rather than using the pseudo-inverse filter h^{-1} of h, a specific deconvolution function $\tilde{h}$ is applied on the reconstructed image to reduce the blur of the optics. To avoid strong aliasing artifacts, this deconvolution function is not the direct inverse of the PSF h but a function such that the deconvolved image would be similar to the ouput of an ideal instrument with the target PSF h_t [Lier 2008]

$$\tilde{h} * h = h_t. \tag{3.12}$$

The idea of using a target PSF h_t is to enforce some specifications on the final image such as the sampling grid and the value of the MTF at the Nyquist frequency. The deconvolution function $\tilde{h}$ is then fully characterized by the target PSF h_t which is mainly obtained from image analysis of empirical results [Lier 2008]. This deconvolution function reduces the blur of the image and enhances the high frequencies of both the image and the noise. The deconvolved image appears thus to be sharp but noisy. The second step of the restoration consists then in a denoising technique on the deconvolved image to reduce the amplified noise. Due to the specific frequential aspect of the deconvolution function h_t, the deconvolved noise is colored, meaning that it occupies a certain band of high frequencies.

State-of-the-art denoising techniques are usually based on the classical wavelet transform which does not have a spectral representation fine enough to capture these bands of high frequencies. For this reason, the denoising technique used by the CNES is based on the method proposed in [Kalifa 2003b] and uses a wavelet packet transform coupled with a (soft-)thresholding of the wavelet coefficients [Lier 2008].

The wavelet packet transform is an extension of the classical wavelet transform and performs iteratively the decomposition on both the low and the high frequencies of the image, contrary to the classical wavelet transform which iterates the decomposition only on the low frequency. As mentioned in Section 2.1.2.1, a wavelet packet transform allows to obtain a finer frequential resolution of the image and to capture specific bands of frequencies. The frequencies bands that are assumed to be noised are then thresholded to reduce the noise power.

To compute the threshold parameters, an image of noise is generated and deconvolved using the deconvolution function $\tilde{h}$. The variance of the deconvolved noise is computed in each subband and compared to the variance of the deconvolved image in the same subband. If these variance are almost the same, then it is assumed that the corresponding subband only contains noise and can be thresholded. The threshold parameter is then computed such that a fixed signal-to-noise ratio is obtained at the output of the restoration. Some reconstruction results of the complete image chain are displayed on Fig. 3.7.

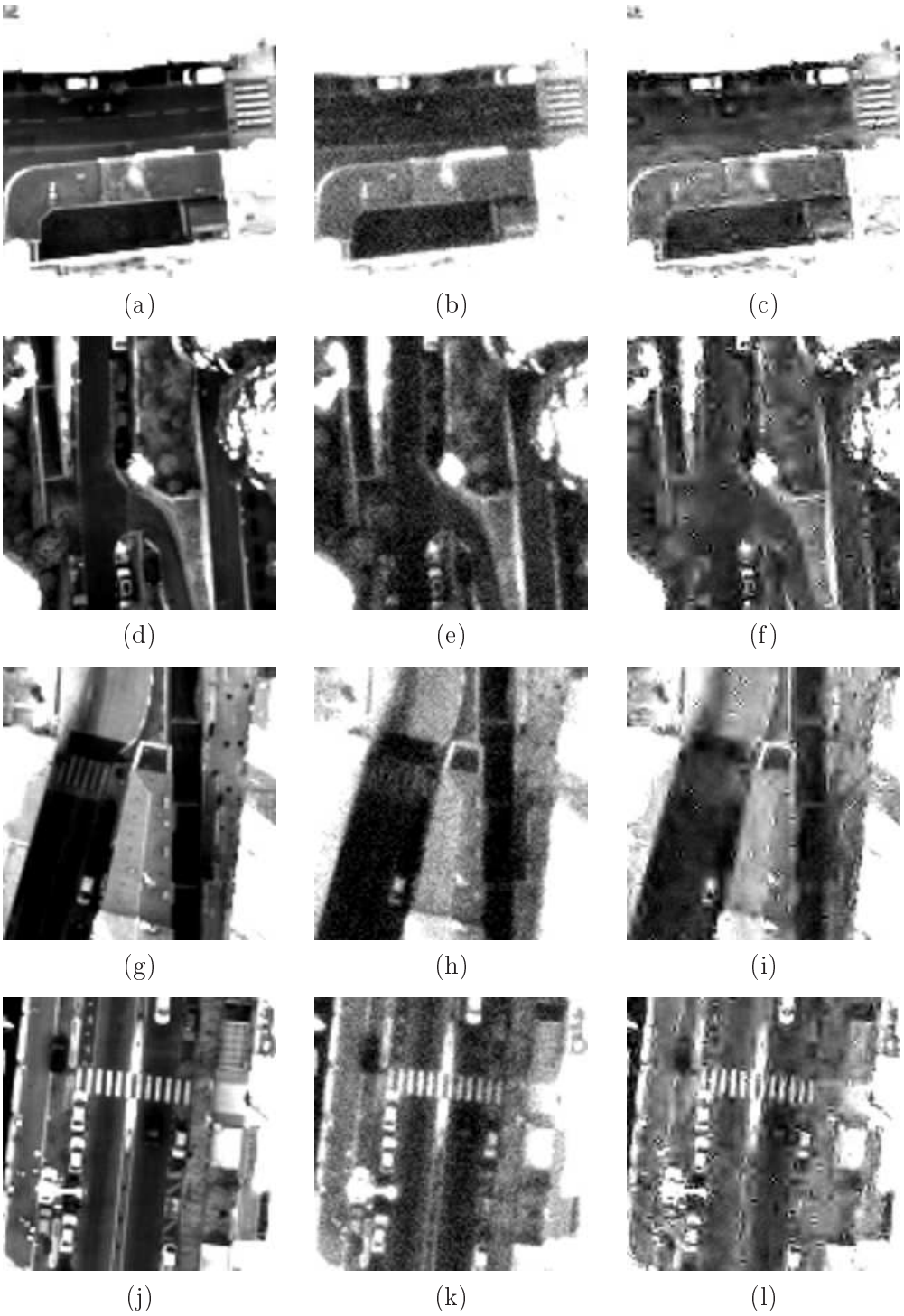


Figure 3.7: Visual result of the imaging chain used by the CNES. Displayed images have a size of 200×200 pixels. For each ligne, the image on the left is a zoom of the clean reference image, the image in the middle is a zoom of the instrumental image, and the image on the right is a zoom of the final image provided by the CNES. The target rate is 2.5 bits/pixel and the simulated SNR is 30-100. The image range has been extended to point up the image reconstruction artifacts.

Part II

Global optimization of the satellite chain

Chapter 4

Optimization of the chain: A theoretical study

In this chapter, we study the theoritical optimization of the global imaging chain. As mentioned in Section 1.1, solving theoritically the global optimization problem (1.1) is a difficult task. Thus, we first reduce the study to the case the image is only degraded by noise and we focus on the optimization of the imaging chain, for three different configurations of coding and restoration, where the global distortion is measured by the mean square error (MSE). We present in Section 4.1 our hypotheses and notations. Section 4.2 is dedicated to the analysis and the optimization of the global distortion for different configurations of the imaging chain. We conclude in Section 4.5 and present perspectives of the study.

4.1 Notations and hypotheses

4.1.1 Notations

For the study, we denote the operators (coding and restoration) applied to the image with a bold uppercase letter. The non-bold uppercase letters represent random variables whose realizations are denoted by a lowercase letter. With this notation, x is a realization of the random variable X. $(X)_i$ denotes the ith element of the random variable X. These variables are multidimensional $x \in \mathbb{R}^N$ where N is the number of pixels. W_x is a random variable associated to the wavelet transform of x and we denote $W_{x,j}$, $j \in \{0, \dots, J-1\}$ (J being the number of subbands) the jth subband of the random variable W_x. A wavelet subband of x is then noted $w_{x,j} \in \mathbb{R}^{N_j}$ where N_j is the size of the subband. Finally, we suppose that a wavelet subband $w_{x,j}$ follows a generalized centered Gaussian distribution law of parameter $\alpha_{w_{x,j}} > 0$ and variance $\sigma^2_{w_{x,j}} > 0$ [Antonini 1992]. The probability density function $p_{w_{x,j}}$ associated to the wavelet subband $w_{x,j}$ can then be modeled as

$$p_{w_{x,j}}(w_{x,j}) = \frac{A\left(\alpha_{w_{x,j}}\right)}{\sigma_{w_{x,j}}} e^{-\left|B\left(\alpha_{w_{x,j}}\right)\frac{w_{x,j}}{\sigma_{w_{x,j}}}\right|^{\alpha_{w_{x,j}}}}, \tag{4.1}$$

with

$$A\left(\alpha_{w_{x,j}}\right) = \frac{\alpha_{w_{x,j}} B\left(\alpha_{w_{x,j}}\right)}{2\Gamma\left(1/\alpha_{w_{x,j}}\right)} \tag{4.2}$$

$$B\left(\alpha_{w_{x,j}}\right) = \sqrt{\frac{\Gamma\left(3/\alpha_{w_{x,j}}\right)}{\Gamma\left(1/\alpha_{w_{x,j}}\right)}}, \tag{4.3}$$

and Γ is the usual Gamma function. The parameters $\sigma^2_{w_{x,j}}$ and $\alpha_{w_{x,j}}$ of the distribution law will be estimated using the kurtosis-based technique proposed in [Kasner 1999]. Note that the same assumption will be applied to all wavelet transforms in the chain with, of course, different distribution parameters.

4.1.2 Coding and denoising operators

As mentioned previously, we study the case the image is only degraded by an instrumental noise z that we assume to be independent, identically distributed and to follow a centered normal distribution with variance σ_z^2. We consider the special case of coding techniques based on wavelet transforms [Shapiro 1993, Said 1996] and [Taubman 2000]. The coding step is then approximately decomposed in a non-redundant wavelet transform followed by a scalar subband quantizer. Note that this approximation is actually close to the coding schemes presented in the cited works.

The wavelet transform is then denoted $\mathbf{W}$ and $\tilde{\mathbf{W}}$ for the inverse transform. Each wavelet subband of the image to encode will be quantized using an infinite mid-tread scalar subband quantizer $\mathbf{Q}$ of step $\Delta_j > 0$ defined as

$$\mathbf{Q}(w_j) = \Delta_j \left\lfloor \frac{w_j}{\Delta_j} + \frac{1}{2} \right\rfloor, \tag{4.4}$$

where $\lfloor\,\rfloor$ is the floor function which returns the greatest integer less than or equal to its argument. Each quantized subband will then be coded using an entropy encoder. Note that the entropy encoding operation does not introduce any degradation in the chain.

For the first part of the study, we also consider that the denoising step is performed in the same wavelet basis than the coding. This choice may however need further explanations. Usually, an efficient wavelet transform for image denoising strongly differs from a wavelet transform suited for image coding. Image denoising techniques actually require redundant wavelet transforms to represent the characteristics of an image such as contours and oriented details while increasing the number of coefficients in image compression may be problematic [Chappelier 2006]. Hence, a non-redundant wavelet transform leads most of the time to poor denoising results. We are however very confident that using the same basis for both coding and denoising may provide a decoding-denoising structure gathered in a single fast and low resources algorithm. Extending the current work to complex denoising schemes such as [Donoho 1995a] is a difficult task that still need to be addressed.

The denoising algorithm $\mathbf{R}$ that we propose to use is then a Tikhonov regularized algorithm which operates independently on the wavelet coefficients of each subband

j of the image. Let $\tilde{w}_j$ be some noisy wavelet subband (of size N_j), its denoised version $\hat{w}_j$ writes

$$\hat{w}_j = \begin{array}{ll} \arg\min & \|w - \tilde{w}_j\|_2^2 + \lambda_j \|w\|_2^2 \ , \\ \text{subject to} & w \in \mathbb{R}^{N_j} \end{array} \tag{4.5}$$

where $\lambda_j > 0$ is a regularizing parameter. The restoration algorithm (4.5) has a closed-form solution which writes

$$\hat{w} = \frac{\tilde{w}}{1 + \lambda_j}. \tag{4.6}$$

We are aware of the simplicity of the considered algorithm, it appears however that the linearity of the restoration algorithm $\mathbf{R}$ is required if one wants to write the global distortion in closed-form. As mentioned previously, much work need to be addressed to consider state-of-the-art denoising algorithms. We now detail the proposed method to perform a global optimization of the global distortion.

4.2 Global optimization of the imaging chain

This section is dedicated to the analysis and the optimization of the global distortion. From Section 3.3.2, we mentioned that, in a general context, the restoration method used by the CNES only deals with the blur and the additive Gaussian noise of the instrument. It actually does not take into account the fact that the transmitted image is also deteriorated with coding noise. The restoration technique used on-ground (i.e. after coding) is therefore also suitable to be used on-board just before coding on the intrumental image, as this image perfectly matches the image formation model considered by the restoration.

From this remark, we decline in this section the theoritical study of the global optimization to the case the restoration is performed before coding or splitted in two parts (one part before coding to reduce the instrumental noise and the other part after coding to process the coding noise).

4.2.1 Optimization of the on-ground chain

4.2.1.1 Presentation of the imaging chain

We first study the on-ground chain where the denoising is performed after coding/decoding, i.e. "on ground". This chain is represented in detail Fig. 4.1. We recall that x is the original image, $\hat{x}$ is the restored one. The instrumental image y is a deteriorated version of the original image x where an additive instrumental noise z has been added. The wavelet subbands of the instrumental image are denoted $w_{y,j}$, $j \in \{0, \ldots, J-1\}$. The quantized and restored version of these subbands are respectively denoted $w_{\tilde{y},j}$ and $w_{\hat{x},j}$.

We further introduce several notations. Let $w_{b,j}$ be the coding error of the subband j

$$w_{b,j} = \mathbf{Q}(w_{y,j}) - w_{y,j}. \tag{4.7}$$

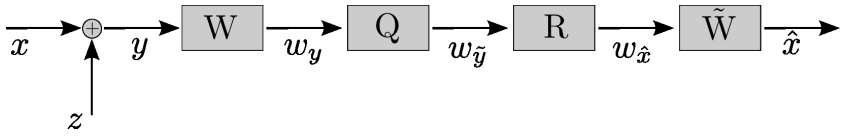


Figure 4.1: Considered on-ground imaging chain

We have

$$\begin{aligned} w_{\tilde{y},j} = \mathbf{Q}(w_{y,j}) &= w_{y,j} + w_{b,j} \\ &= w_{x,j} + w_{z,j} + w_{b,j} \\ &= w_{x,j} + w_{\varepsilon,j}, \end{aligned} \tag{4.8}$$

where $w_{\varepsilon,j} = w_{z,j} + w_{b,j}$ is referred hereafter to the global error. The main hypothesis of the proposed method is to consider the first-order moments of the term $w_{\varepsilon,j}$ to be independent to the ones of $w_{x,j}$, that is

$$E\left[W_{\varepsilon,j}^m W_{x,j}^n\right] = E\left[W_{\varepsilon,j}^m\right] E\left[W_{x,j}^n\right] \tag{4.9}$$

for any integer $m > 0$, $n > 0$ and where $W_{\varepsilon,j}$ and $W_{x,j}$ are the random variables associated to $w_{\varepsilon,j}$ and $w_{x,j}$. This hypothesis is mainly based on the fact that the quantizing part of the scheme Fig. 4.1 can be seen as a non-substractive dithering system where the Gaussian instrumental noise z acts as a dithering noise [Wannamaker 2000].

We detail in the next part this hypothesis of decorrelation.

4.2.1.2 Decorrelation hypothesis

A dithering system consists in inserting a noise with a certain probability density function prior to quantizing, to improve the decorrelation property [Vanderkooy 1987]. As mentioned in [Wannamaker 2000], a non-substractive dithering system (named non-substractive as the dithering noise is not substracted after quantizing) allows the moments of the global error (that is the sum of the coding error and dithering noise) to be fully decorrelated to the moments of the coding source.

It happens that a Gaussian distribution, if its standard deviation is large enough [Vanderkooy 1987], stands among the probability density functions which allow a noise to be considered as a dithering noise. The idea here is then to take benefit of the presence of the instrumental noise by considering it as a dithering noise. With such consideration, we know that the m first-order moments of the global error are decorrelated to the n first-order moments of the quantizing source, giving property (4.9).

Moreover, if the instrumental noise z meets the dithering noise requirements, we

also have [Wannamaker 2000]

$$E\left[W_{\varepsilon,j}\right] = 0, \tag{4.10}$$

$$E\left[\|W_{\varepsilon,j}\|^2\right] = N_j \sigma^2_{w_{z,j}} + N_j \frac{\Delta_j^2}{12}, \tag{4.11}$$

where $\sigma_{w_{z,j}}$ is the standard deviation of the distribution law of the wavelet transform $w_{z,j}$. A more developed presentation of dithering techniques is included in Appendix B. The standard deviation required by a Gaussian noise to effectively acts as a dithering noise has been studied in [Vanderkooy 1987]. In the present case, the condition (4.9) will be verified if the following statement is true

$$\sigma_{w_{z,j}} > \frac{\Delta_j}{2}. \tag{4.12}$$

As the standard deviation of instrumental noise is usually low in imaging systems, the condition (4.12) assumes that the proposed approach will be valid only for high coding rates. We will however develop our method to consider all coding rates.

4.2.1.3 Analysis of the global distortion

As mentioned in the Section 4.1, the studied imaging chain depends on two sets of parameters: The denoising parameters λ_j in (4.6) and the quantizing steps Δ_j in (4.4), for each $j \in \{0, \ldots, J-1\}$. The global coding/denoising joint optimization problem consists in finding the sets $\{\lambda_j^*\}$ and $\{\Delta_j^*\}$ of optimal parameters which minimize, on average, the global distortion D under the constraint that the coding rate R does not exceed the target rate R_c. This global rate-distortion-denoising joint optimization problem can be formalized as the following

$$\begin{aligned} \{\lambda_j^*\}, \{\Delta_j^*\} = \quad & \underset{}{\arg\min} \quad D\left(\{\lambda_j\}, \{\Delta_j\}\right) \\ & \text{subject to} \quad R\left(\{\lambda_j\}, \{\Delta_j\}\right) \leq R_c, \\ & \qquad\qquad \lambda_j > 0, \forall j \in \{0, \ldots, J-1\} \\ & \qquad\qquad \Delta_j > 0, \forall j \in \{0, \ldots, J-1\} \end{aligned} \quad . \tag{4.13}$$

Under this form, the optimization problem (4.13) is difficult to solve so that it is usually written under an unconstrained form [Everett 1963]. Let $\tau > 0$ be a Lagrange multiplier. The Lagrange dual function L writes

$$\begin{aligned} L(\tau) = \quad & \inf \quad D(\{\lambda_j\}, \{\Delta_j\}) + \tau \left(R(\{\lambda_j\}, \{\Delta_j\}) - R_c\right) \\ & \qquad \lambda_j > 0,\ j \in \{0, \ldots, J-1\} \\ & \qquad \Delta_j > 0,\ j \in \{0, \ldots, J-1\} \end{aligned} \quad . \tag{4.14}$$

Problem (4.13) can then be written [Boyd 2004]

$$\{\lambda_j^*\}, \{\Delta_j^*\} = \max_{\tau > 0} L(\tau). \tag{4.15}$$

To solve the global distortion joint optimization problem (4.15), we need to express the mean global distortion D and the global coding rate R as a function of the sets of regularizing parameters $\{\lambda_j\}$ and quantizing steps $\{\Delta_j\}$.

Proposition 1. *If $\sigma_{w_{z,j}}$ verifies hypothesis (4.12) for each $j \in \{0, \dots, J-1\}$, then the mean global distortion D of the imaging chain displayed Fig. 4.1 writes*

$$D(\{\lambda_j\}, \{\Delta_j\}) = \sum_{j=0}^{J-1} \frac{\pi_j a_j \lambda_j^2}{(1+\lambda_j)^2} \sigma^2_{w_{x,j}} + \frac{\pi_j a_j}{(1+\lambda_j)^2} \sigma^2_{w_{z,j}} + \frac{\pi_j a_j}{(1+\lambda_j)^2} \frac{\Delta_j^2}{12}, \tag{4.16}$$

where

$$a_j = \frac{N_j}{N}, \tag{4.17}$$

is the weight of the subband j in the whole image.

Proof. We start from the fact that the mean global distortion writes

$$D(\{\lambda_j\}, \{\Delta_j\}) = \frac{1}{N} E\left(\|X - \hat{X}\|^2\right), \tag{4.18}$$

where $\hat{X}$ is the random variable associated to the output final image $\hat{x}$. Thanks to the orthogonality of the wavelet subbands, the global distortion can also be formulated as

$$D(\{\lambda_j\}, \{\Delta_j\}) = \frac{1}{N} \sum_{j=0}^{J-1} \pi_j E\left(\|W_{x,j} - W_{\hat{x},j}\|^2\right), \tag{4.19}$$

where π_j are weighting coefficients which depend on the filters and the decimation factors used in the wavelet transform [Usevitch 1996]. Note that these weighting coefficients are only required if one considers biorthogonal wavelet transforms such as the CDF 9/7 wavelet transform [Cohen 1992]. They are equal to 1 for an orthogonal wavelet transform.

In the case of the studied imaging chain displayed Fig. 4.1, the final image is the output of the restoration and writes

$$w_{\hat{x},j} = \mathbf{R} w_{\tilde{y},j}. \tag{4.20}$$

Using (4.6) and (4.8), the final image can be expressed as a function of the source and the global error

$$w_{\hat{x},j} = \frac{w_{x,j}}{1+\lambda_j} + \frac{w_{\varepsilon,j}}{1+\lambda_j}. \tag{4.21}$$

From (4.19), (4.21) and using the moments decorrelation hypothesis (4.9), we deduce the global distortion

$$\begin{aligned} D(\{\lambda_j\}, \{\Delta_j\}) &= \frac{1}{N} E\left(\|X - \hat{X}\|^2\right) \\ &= \frac{1}{N} \sum_{j=0}^{J-1} \frac{\pi_j \lambda_j^2}{(1+\lambda_j)^2} E\left(\|W_{x,j}\|^2\right) + \frac{\pi_j}{(1+\lambda_j)^2} E\left(\|W_{\varepsilon,j}\|^2\right). \end{aligned} \tag{4.22}$$

Finally, the global distortion (4.22) can be further developed using the results (4.11) to obtain the expression (4.16). □

Note that the global distortion (4.16) requires the knowledge of the variance of each subband of the original image $\sigma^2_{w_{x,j}}$. This variance is generally unknown but can be roughly deduced from the observed image. For an orthogonal or a biorthogonal wavelet transform, the variance of the noise in each wavelet subband j is equal (or almost equal in the case of a biorthogonal wavelet transform) to the variance of the noise in the image domain, i.e. $\sigma^2_{w_{z,j}} = \sigma^2_z$, where σ_z is supposed to be known. Then, $\sigma^2_{w_{x,j}}$ can be approximately computed during the rate-allocation of the coder from the observed subband variance $\sigma^2_{w_{y,j}}$ by

$$\sigma^2_{w_{x,j}} = \sigma^2_{w_{y,j}} - \sigma^2_z. \tag{4.23}$$

The second part of the problem (4.15) requires the expression of the global coding rate R. This rate can be expressed as the weighted sum of the rate in each subband R_j

$$R(\{\lambda_j\}, \{\Delta_j\}) = \sum_{j=0}^{J-1} a_j R_j(\Delta_j), \tag{4.24}$$

where a_j is given in (4.17). As mentioned in Section 4.1.2, we assume that each quantized subband is encoded using an entropy encoder. The coding rate R_j of a subband j can then be estimated by its entropy [Shannon 1948]

$$R_j(\Delta_j) = -\sum_{m=-\infty}^{+\infty} P_{w_{y,j}}(m, \Delta_j) \log_2 \left(P_{w_{y,j}}(m, \Delta_j) \right), \tag{4.25}$$

where $P_{w_{y,j}}(m, \Delta_j)$ is the probability to get the symbol m which depends on the density probability function $p_{w_{y,j}}$ of the subband $w_{y,j}$ and on the quantizing step Δ_j

$$P_{w_{y,j}}(m, \Delta_j) = \int_{m\Delta_j - \frac{\Delta_j}{2}}^{m\Delta_j + \frac{\Delta_j}{2}} p_{w_{y,j}}(w_{y,j}) dw_{y,j}. \tag{4.26}$$

From Section 4.1.1, we assume that each wavelet subband follows the generalized centered Gaussian distribution law defined in (4.1). The density probability function $p_{w_{y,j}}$ is then given by

$$p_{w_{y,j}}(w_{y,j}) = \frac{A\left(\alpha_{w_{y,j}}\right)}{\sigma_{w_{y,j}}} e^{-\left| B(\alpha_{w_{y,j}}) \frac{w_{y,j}}{\sigma_{w_{y,j}}} \right|^{\alpha_{w_{y,j}}}}, \tag{4.27}$$

with

$$A\left(\alpha_{w_{y,j}}\right) = \frac{\alpha_{w_{y,j}} B\left(\alpha_{w_{y,j}}\right)}{2\Gamma\left(1/\alpha_{w_{y,j}}\right)} \tag{4.28}$$

$$B\left(\alpha_{w_{y,j}}\right) = \sqrt{\frac{\Gamma\left(3/\alpha_{w_{y,j}}\right)}{\Gamma\left(1/\alpha_{w_{y,j}}\right)}}, \tag{4.29}$$

and where $\sigma^2_{w_{y,j}}$ and $\alpha_{w_{y,j}}$ are the parameters of the distribution law, estimated using the kurtosis-based technique proposed in [Kasner 1999]

$$\sigma^2_{w_{y,j}} = E[w^2_{y,j}], \tag{4.30}$$

$$\alpha_{w_{y,j}} = \frac{1.447}{\log\left(\frac{E[w^4_{y,j}]}{E[w^2_{y,j}]^2}\right) - 0.345}. \tag{4.31}$$

Proposition 2. *The global rate-distortion optimization problem (4.13) can be solved by maximizing*

$$L(\tau) = \begin{array}{ll} \inf & \phi_\tau(\{\Delta_j\},\{\lambda_j\}) \\ & \lambda_j > 0,\ j \in \{0,\dots,J-1\} \\ & \Delta_j > 0,\ j \in \{0,\dots,J-1\} \end{array}, \tag{4.32}$$

with respect to $\tau > 0$ and where

$$\phi_\tau(\{\Delta_j\},\{\lambda_j\}) = \sum_{j=0}^{J-1} \frac{\pi_j a_j \lambda_j^2}{(1+\lambda_j)^2}\sigma^2_{w_{x,j}} + \frac{\pi_j a_j}{(1+\lambda_j)^2}\sigma_z^2 + \frac{\pi_j a_j \Delta_j^2}{12(1+\lambda_j)^2} + \tau\left(\sum_{j=0}^{J-1} a_j R_j(\Delta_j) - R_c\right). \tag{4.33}$$

Proof. This demonstration is straightforward. From (4.15), we define

$$\phi_\tau(\{\Delta_j\},\{\lambda_j\}) = D(\{\Delta_j\},\{\lambda_j\}) + \tau\left(R(\{\Delta_j\},\{\lambda_j\}) - R_c\right), \tag{4.34}$$

and we substitute D and R with their respective expressions (4.16) and (4.24). We further simplify (4.16) using the approximation $\sigma^2_{w_{z,j}} = \sigma_z^2$. The reformulation of problem (4.13) is then obtained using (4.14) and (4.15). □

We detail in the next part how to solve problem (4.13).

4.2.1.4 Global rate-distortion-denoising optimization

Using proposition 2, the optimization problem (4.13) becomes

$$\{\Delta_j^*\},\{\lambda_j^*\} = \max_{\tau>0}\left(\begin{array}{ll} \inf & \phi_\tau(\{\Delta_j\},\{\lambda_j\}) \\ & \lambda_j > 0,\ \forall j \in \{0,\dots,J-1\} \\ & \Delta_j > 0,\ \forall j \in \{0,\dots,J-1\} \end{array}\right). \tag{4.35}$$

The existence and uniqueness of solutions of problem (4.35) is not straightforward but we can show that a solution of problem (4.35) exists and is unique (see Appendix A.2). We propose a numerical algorithm to find this solution. This algorithm is based on the resolution of the simultaneous equations obtained from the KKT conditions [Kuhn 1951] of problem (4.35).

Proposition 3. *The KKT conditions of problem (4.35) admits only one solution* $(\{\lambda_j^*\}, \{\Delta_j^*\}, \tau^*)$ *which verifies*

$$\lambda_j^* = \frac{\sigma_z^2}{\sigma_{w_{x,j}}^2} + \frac{\Delta_j^{*2}}{12\sigma_{w_{x,j}}^2}, \quad \forall j \in \{0, \dots, J-1\} \tag{4.36}$$

$$\frac{\pi_j \Delta_j^*}{6(1+\lambda_j)^2} + \tau^* \frac{\partial R_j}{\partial \Delta_j}(\Delta_j^*) = 0, \quad \forall j \in \{0, \dots, J-1\} \tag{4.37}$$

$$\sum_{j=0}^{J-1} a_j R_j(\Delta_j^*) = R_c. \tag{4.38}$$

Proof. From the KKT conditions of problem (4.35), we get (see Appendix A.2)

$$\frac{\partial \phi(\Delta_j^*, \lambda_j^*, \tau^*)}{\partial \Delta_j} = \frac{a_j \pi_j \Delta_j^*}{6(1+\lambda_j^*)^2} + \tau^* a_j \frac{\partial R_j}{\partial \Delta_j}(\Delta_j^*) = 0 \tag{4.39}$$

$$\frac{\partial \phi(\Delta_j^*, \lambda_j^*, \tau^*)}{\partial \tau} = \sum_{j=0}^{J-1} a_j R_j(\Delta_j^*) - R_c = 0 \tag{4.40}$$

$$\frac{\partial \phi(\Delta_j^*, \lambda_j^*, \tau^*)}{\partial \lambda_j} = \frac{12 a_j \pi_j \lambda_j^* \sigma_{w_{x,j}}^2 - 12 a_j \pi_j \sigma_z^2 - a_j \pi_j \Delta_j^{*2}}{6(1+\lambda_j^*)^3} = 0 \tag{4.41}$$

$$\tag{4.42}$$

with

$$\frac{\partial R_j}{\partial \Delta_j}(\Delta_j) = -\frac{1}{\log(2)} \sum_{m=-\infty}^{+\infty} \left[1 + \log\left(P_{w_{y,j}}(m, \Delta_j)\right)\right] \times$$
$$\left[p_{w_{y,j}}\left(m\Delta_j + \frac{\Delta_j}{2}\right)\left(m + \frac{1}{2}\right) - p_{w_{y,j}}\left(m\Delta_j - \frac{\Delta_j}{2}\right)\left(m - \frac{1}{2}\right)\right]. \tag{4.43}$$

The expression (4.36) and conditions (4.37) and (4.38) on the optimal parameters directly follow from the optimality conditions (4.39). The existence and uniqueness of these parameters is much longer and is addressed in Appendix A.2. □

As we can see from (4.36), (4.37) and (4.38), the parameters $\{\Delta_j^*\}$ and τ^* can not be computed analytically. But as mentioned in Appendix A.2, any root-finding algorithms can be used to achieve this goal. For our simulations, binary search algorithms will be used for the computation of both $\{\Delta_j^*\}$, τ^* and for the sake of simplicity, each binary search algorithm will be parametrized to the same given precision $\rho = 0.1$.

The case of the low frequency subband ($j = J-1$) will be processed differently as we do not want to degrade these coefficients. We will only use quantizing to round these coefficients to their nearest integers. Consequently, we will set

$$\Delta_{J-1}^* = 1, \tag{4.44}$$

$$\lambda_{J-1}^* = \frac{\sigma_z^2}{\sigma_{w_{x,J-1}}^2} + \frac{1}{12\sigma_{w_{x,J-1}}^2}. \tag{4.45}$$

Finally, the overall joint optimization procedure for solving problem (4.13) is given in the Algorithm 1. Note that the binary search sub-procedures are not detailled in this process. The Algorithm 1 intends to be quite general and we let the choice of the root-finding algorithms to the user.

Algorithm 1 Global rate-distortion-denoising joint optimization algorithm for the on-ground imaging chain

Set $\tau = 1$.
Set $\rho = 0.1$.
while $\left|\sum_{j=0}^{J-1} a_j R_j - R_c\right| > \rho$ **do**
  for j from 0 to $J-2$ **do**
    Set $\Delta_j = 1$.
    Compute the value of the regularizing parameter λ_j from (4.36).
    while $\left|\frac{\pi_j \Delta_j}{6(1+\lambda_j)^2} + \tau \frac{\partial R_j}{\partial \Delta_j}(\Delta_j)\right| > \rho$ **do**
      Increase the value of Δ_j.
      Compute the value of the regularizing parameter λ_j from (4.36).
    end while
  end for
  Set $\Delta_{J-1} = 1$.
  Compute the regularizing paramater λ_{J-1} from (4.45).
  if $\left|\sum_{j=0}^{J-1} a_j R_j - R_c\right| > \rho$ **then**
    Increase the value of τ.
  end if
end while
Output the optimal regularizing parameters $\{\lambda_j^*\}$.
Output the optimal quantizing steps $\{\Delta_j^*\}$.

4.2.1.5 Results

We simulate the joint optimization Algorithm 1 on the high-dynamic range remote sensing image displayed Fig. 4.2. For this simulation, we set the wavelet transform $\mathbf{W}$ to be a three levels CDF 9/7 wavelet transform [Cohen 1992] and the restoration $\mathbf{R}$ is given by (4.6). The image has been noised with an additive white Gaussian noise with different standard deviations σ_z, as the efficiency of the proposed estimation depends on σ_z, see Eq. (4.12). The following cases have been tested $\sigma_z \in \{25, 50, 75, 100\}$.

For each target rate, we simulate the imaging chain given Fig. 4.1 with the usual disjoint optimization technique, which consists in selecting the quantizing steps and the regularizing parameters such that the coding and the restoration errors are independently minimized. The coding error minimization has been achieved using the rate-distortion allocation based model proposed in [Parisot 2001]. As for the restoration error, it has been minimized using an exhaustive search of the optimal

Figure 4.2: Reference image, Cannes harbour (12 bits panchromatic image, 30 cm resolution, 1024×1024 pixels).

regularizing parameters. Once the final image has been reconstructed using these parameters, we numerically compute the global distortion

$$D = \frac{1}{N}\|x - \hat{x}\|^2, \tag{4.46}$$

where x is the clean (i.e. noiseless) test image, assumed to be known in our numerical experiments, and $\hat{x}$ is the final image. The distortion (4.46) is the true distortion and will be referred as the ground truth in our simulations. The estimation model (4.16) of the global distortion that we proposed has then been computed with the values of parameters obtained for the ground truth. This allows to verify that the estimation (4.16) of the global distortion is close to the ground truth (4.46), implying the validity of the proposed method. And finally, we use the proposed joint optimization Algorithm 1 to compute the optimal parameters, that we inserted into the estimation model (4.16) to estimate the minimal distortion.

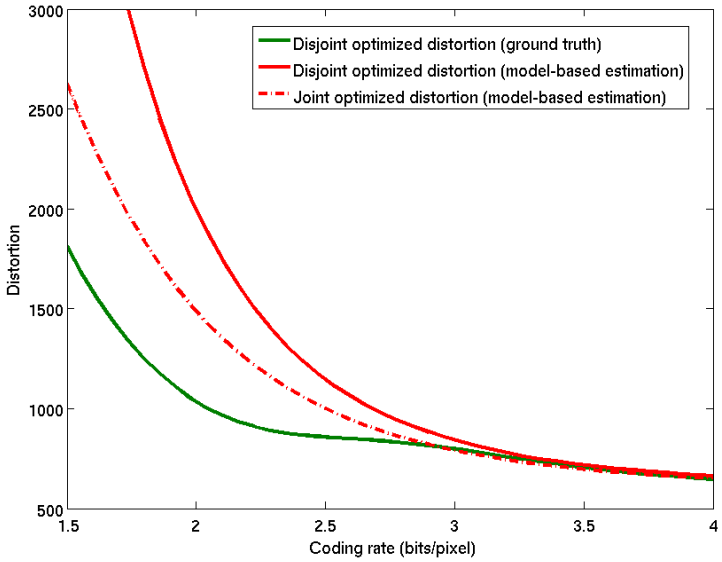


Figure 4.3: Comparison of the disjoint optimized distortion (ground truth and model-based estimation) to the joint optimized distortion (model-based estimation), $\sigma_z = 25$.

Results are given Fig. 4.3 to 4.6. We immediately see that the validity of the proposed estimation, as expected by the hypothesis (4.12), is not always verified and depends on the target coding rate, for a given σ_z. As expected, the proposed estimation approximates well the true distortion, on the simulated cases, for medium to high coding rates but does not give satisfying results for low coding rates. This can be explained by the fact that low target coding rates increase the subbands quantizing steps. Consequently, the condition (4.12) is not respected anymore and the moments of the global error cannot be considered decorrelated to the moments of the source.

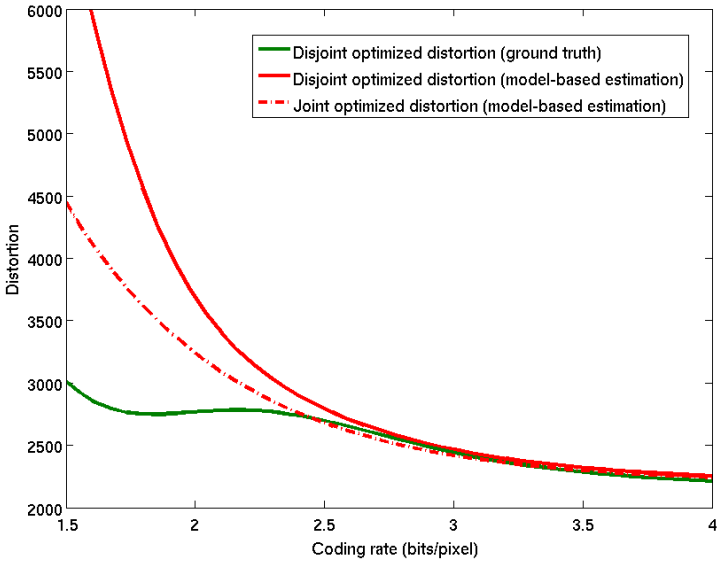


Figure 4.4: Comparison of the disjoint optimized distortion (ground truth and model-based estimation) to the joint optimized distortion (model-based estimation), $\sigma_z = 50$.

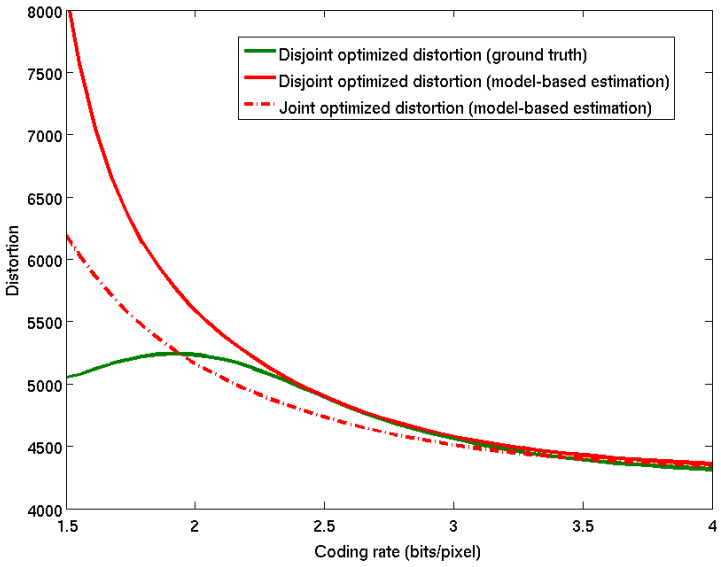


Figure 4.5: Comparison of the disjoint optimized distortion (ground truth and model-based estimation) to the joint optimized distortion (model-based estimation), $\sigma_z = 75$.

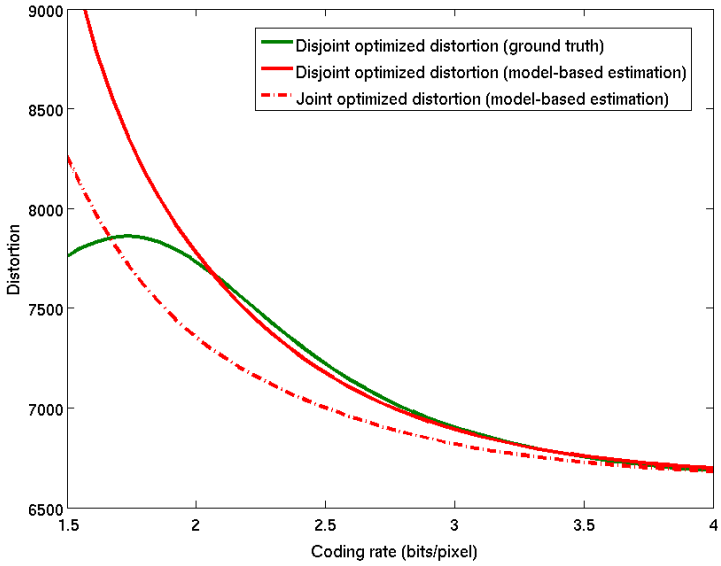


Figure 4.6: Comparison of the disjoint optimized distortion (ground truth and model-based estimation) to the joint optimized distortion (model-based estimation), $\sigma_z = 100$.

To analyse more precisely the range of validity of the proposed estimation, we compute the error (in absolute value) between the ground truth distortion and its model-based estimation (4.16) for the simulated values of standard deviation σ_z. The resulting curve is displayed Fig. 4.7. When the standard deviation is low ($\sigma_z = 25$), we see that the proposed estimation is performant if the coding rate is around 3.5 bits/pixel and more. However for this high coding rate, the coding step is almost lossless such that the global optimization problem is reduced to the optimization of the restoration only. Therefore, the joint and the disjoint optimization techniques become the same and give then similar results.

But the range of validity of the proposed estimation increases as the standard deviation increases. For a high standard deviation ($\sigma_z = 100$), we can verify that the proposed estimation is valid for lower coding rates (around 2.2 bits/pixel and more). In that case, the joint optimization displays significant improvement in comparison to the disjoint optimization. It allows for example to reach the same global error than the disjoint optimized technique but for a lower coding rate. For $\sigma_z = 100$ (Fig. 4.6), the joint optimization technique reaches at 1.73 bits/pixel the same distortion than the one obtained at 2.04 bits/pixels for the disjoint optimization technique, saving therefore 15% of the bit budget. The benefit in term of compression performances of the joint optimization technique appears then to be very significant. This simulated case is however slightly excessive in the case of satellite imaging as the standard deviation of the instrumental noise in a satellite chain is low and rarely exceeds ten on average.

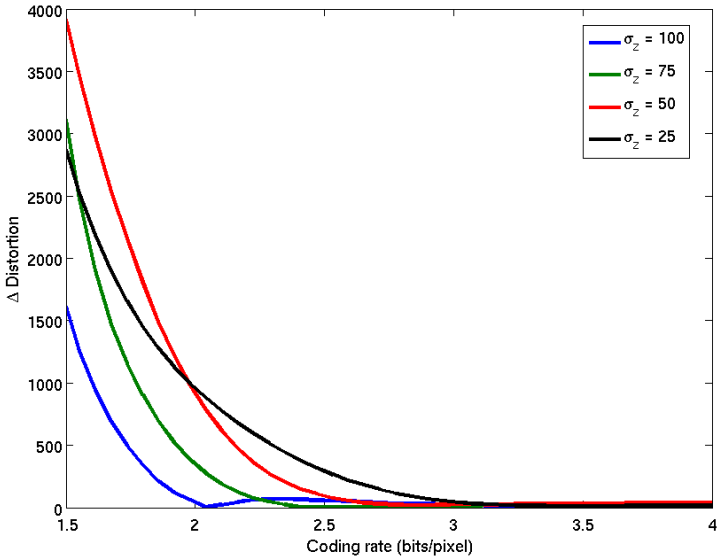
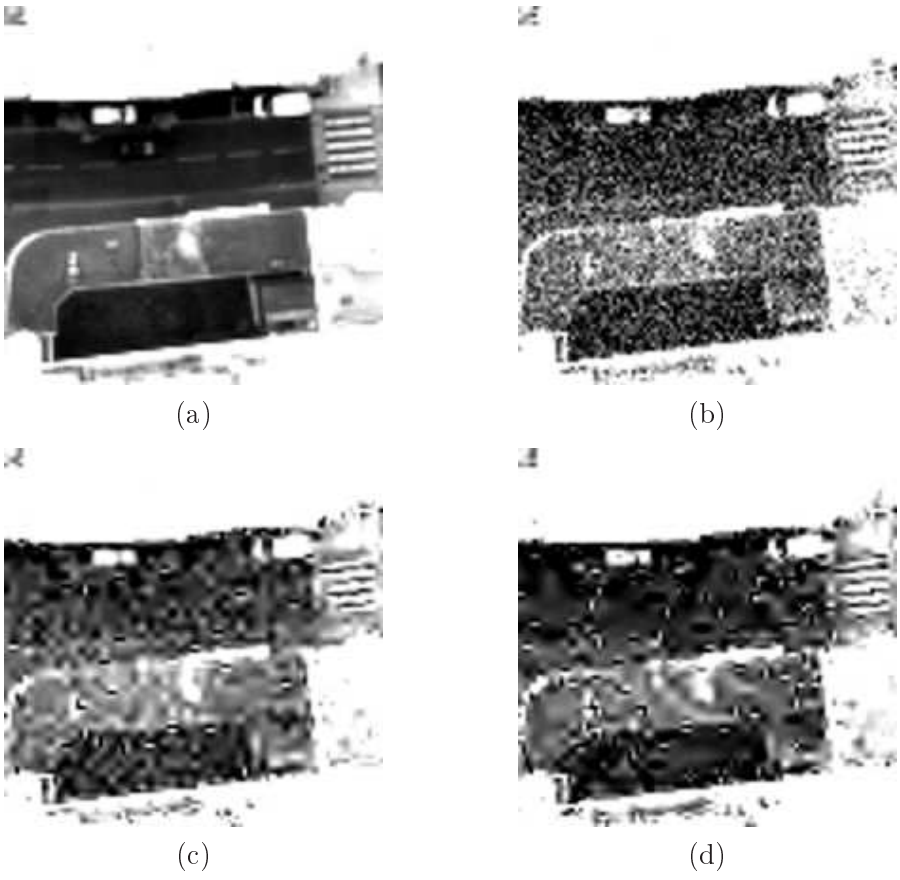


Figure 4.7: Difference (in absolute value) between the ground truth and model-based estimation distortion for the simulated standard deviations of the instrumental noise.

To fit the characterics of a true imaging chain, we simulate the case $\sigma_z = 10$ which is much more representative of the SNR obtained in satellite imaging (see Table 3.1, Page 35). We do not display the rate-distortion curve of this simulation as, similarly to the case $\sigma_z = 25$ displayed Fig. 4.3, the joint and the disjoint optimization techniques are equal in term of distortion. Visual results however differ as shown by Fig. 4.8 to 4.11. We do not focus on the quality of the reconstructed images regarding to the reference one as the considered chain is excessively simple. Clearly, the presence of artifacts on the reconstructed image is due to the simple hypothesis that we made on the restoration algorithm, see Eq. (4.5). On the contrary, we are more concerned on the improvement of the image quality of the joint optimized chain with respect to the disjoint optimized one. We can see that the global joint optimization of the chain always leads to a reconstructed image which contains less blurry edges or ringing artifacts. This is particularly visible on the edges of the buildings Fig. 4.8 and 4.10. It is important to note that the presented visual results have been simulated at a coding rate of 2.5 bits/pixel. And we know that the estimation of the global distortion is not valid at this rate, leading to suboptimal computed parameters. A finer estimation of the global distortion will therefore give better results that the ones displayed here.

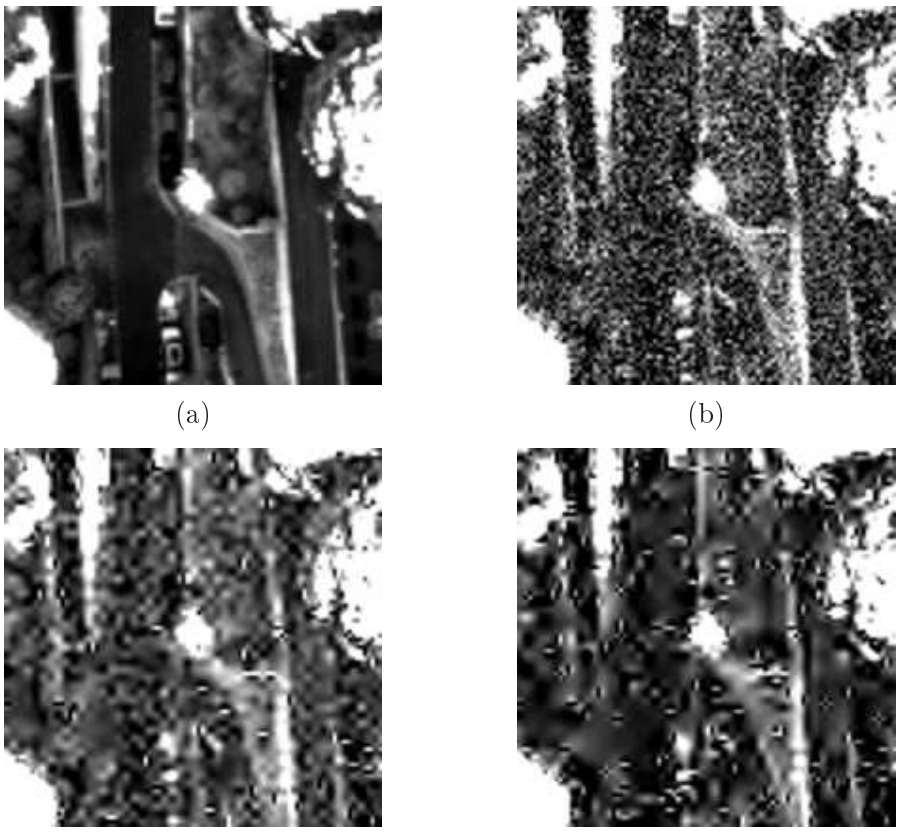
Finally, we see that the obtained results clearly point that optimizing coding and denoising separately is suboptimal. One needs instead to address the problem of imaging chain design in its globality; the proposed method and the obtained results are encouraging in this sense. Extending the proposed method to lower coding rates and to more complex denoising schemes appears however to be difficult to address.



(a) (b)

(c) (d)

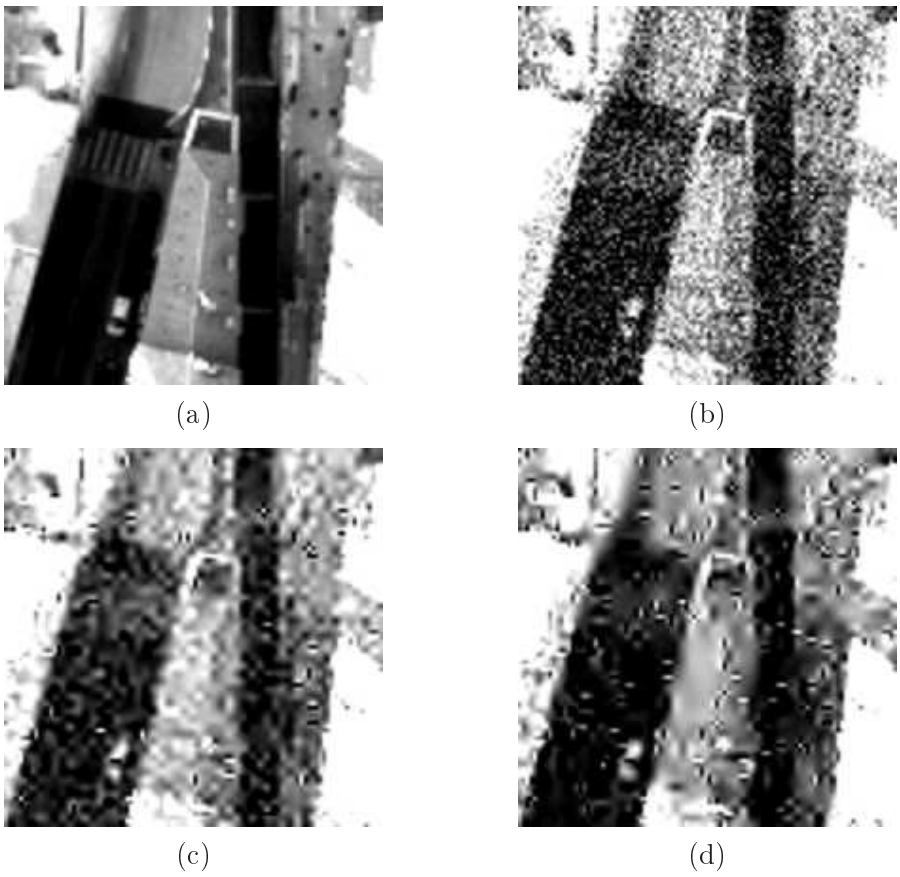
Figure 4.8: Visual comparison of reconstruction results. Displayed images have a size of 200×200 pixels. (a) is the reference image, (b) is the noisy observed image, (c) is the image reconstructed with the parameters obtained by the disjoint minimization of the ground truth distortion and (d) is the image reconstructed with the parameters obtained by the joint optimization, performed using Algorithm 1, of the model-based estimated distortion. The coding rate is 2.5 bits/pixel. The image range has been extended to point up the image reconstruction artifacts.



(a) (b)

(c) (d)

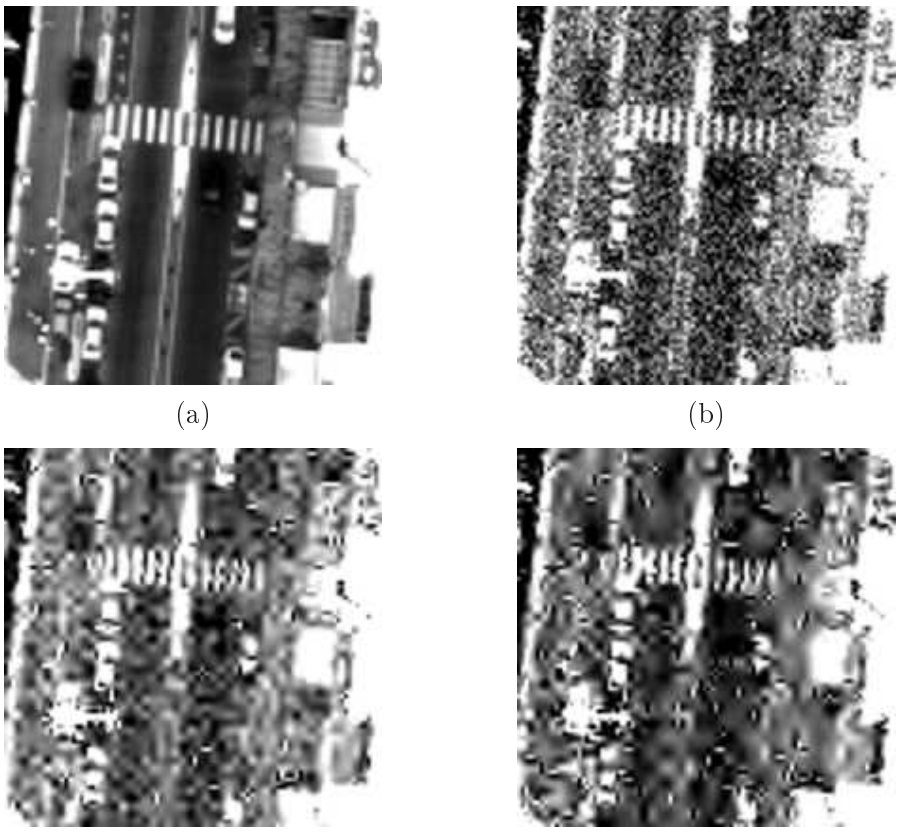
Figure 4.9: Visual comparison of reconstruction results. Displayed images have a size of 200×200 pixels. (a) is the reference image, (b) is the noisy observed image, (c) is the image reconstructed with the parameters obtained by the disjoint minimization of the ground truth distortion and (d) is the image reconstructed with the parameters obtained by the joint optimization, performed using Algorithm 1, of the model-based estimated distortion. The coding rate is 2.5 bits/pixel. The image range has been extended to point up the image reconstruction artifacts.



(a) (b)

(c) (d)

Figure 4.10: Visual comparison of reconstruction results. Displayed images have a size of 200×200 pixels. (a) is the reference image, (b) is the noisy observed image, (c) is the image reconstructed with the parameters obtained by the disjoint minimization of the ground truth distortion and (d) is the image reconstructed with the parameters obtained by the joint optimization, performed using Algorithm 1, of the model-based estimated distortion. The coding rate is 2.5 bits/pixel. The image range has been extended to point up the image reconstruction artifacts.



(a) (b)

(c) (d)

Figure 4.11: Visual comparison of reconstruction results. Displayed images have a size of 200×200 pixels. (a) is the reference image, (b) is the noisy observed image, (c) is the image reconstructed with the parameters obtained by the disjoint minimization of the ground truth distortion and (d) is the image reconstructed with the parameters obtained by the joint optimization, performed using Algorithm 1, of the model-based estimated distortion. The coding rate is 2.5 bits/pixel. The image range has been extended to point up the image reconstruction artifacts.

For this reason, we will propose in Chapter 5 an alternative technique to perform the global optimization.

4.2.2 Optimization of the on-board chain

4.2.2.1 Presentation of the imaging chain

As mentioned in the beginning of Section 4.2, we also studied the imaging chain in the case the denoising is performed before coding, as illustrated on Fig. 4.12. For this imaging chain, the transmitted image is the denoised one and the final image is the one obtained after decoding (we will discuss in Section 4.2.3 the necessity of using a second denoising step after decoding).

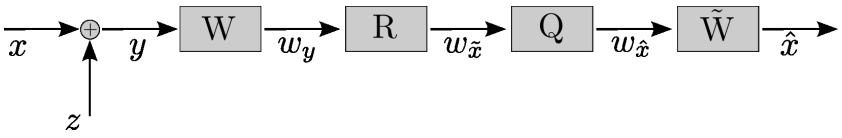


Figure 4.12: Considered on-board imaging chain

Similarly to the chain presented in Section 4.2.1, the instrumental image y is a deteriorated version of the original image x where an additive instrumental noise z has been added. The wavelet subbands of the instrumental image are again denoted $w_{y,j}$, $j \in \{0, \dots, J-1\}$. The restored and quantized version of these subbands are respectively denoted $w_{\tilde{x},j}$ and $w_{\hat{x},j}$.

Let $w_{b,j}$ be the coding error of the subband j

$$w_{b,j} = \mathbf{Q}(w_{\tilde{x},j}) - w_{\tilde{x},j}. \tag{4.47}$$

We have

$$\begin{aligned} w_{\hat{x},j} = \mathbf{Q}(w_{\tilde{x},j}) &= w_{\tilde{x},j} + w_{b,j} \\ &= \frac{w_{y,j}}{1+\lambda_j} + w_{b,j} \\ &= \frac{w_{x,j}}{1+\lambda_j} + \frac{w_{z,j}}{1+\lambda_j} + w_{b,j} \\ &= \frac{w_{x,j}}{1+\lambda_j} + w_{\varepsilon,j}, \end{aligned} \tag{4.48}$$

where $w_{\varepsilon,j} = \frac{w_{z,j}}{1+\lambda_j} + w_{b,j}$ is the global error. We detail in the next part how to formulate an expression of the global distortion.

4.2.2.2 Decorrelation hypothesis

The decorrelation hypothesis (4.9) will also be used to compute the global distortion of the imaging chain presented Fig. 4.12. The main difference is that the quantized image is now the restored one. As a consequence of this restoration, the standard

deviation of the instrumental noise is divided by a factor $1+\lambda_j$, see Eq. (4.48). We have

$$\sigma'_{w_{z,j}} = \frac{\sigma_{w_{z,j}}}{1+\lambda_j}, \tag{4.49}$$

where $\sigma'_{w_{z,j}}$ is the standard deviation of the residual instrumental noise. We know from (4.12) that the decorrelation hypothesis (4.9) is valid only if the standard deviation of the noise presented at the input of the quantizer is greater than half of the quantizing step, i.e.

$$\sigma'_{w_{z,j}} > \frac{\Delta_j}{2}. \tag{4.50}$$

From (4.49) and (4.50), the condition (4.9) will now be verified if the following statement is true

$$\sigma_{w_{z,j}} > \frac{\Delta_j}{2}(1+\lambda_j). \tag{4.51}$$

In comparison to the on-ground imaging chain studied in Section 4.2.1, we see that a factor $(1+\lambda_j)$ has been introduced in the condition (4.51). As $\lambda_j > 0, \forall j \in \{0,\ldots,J-1\}$, the decorrelation hypothesis (4.9) may then be more difficult to verify in the case of the on-board imaging chain. If the instrumental noise z meets the dithering noise requirements, we also have [Wannamaker 2000]

$$E\left[W_{\varepsilon,j}\right] = 0, \tag{4.52}$$

$$E\left[\|W_{\varepsilon,j}\|^2\right] = N_j \frac{\sigma^2_{w_{z,j}}}{(1+\lambda_j)^2} + N_j \frac{\Delta_j^2}{12}. \tag{4.53}$$

4.2.2.3 Analysis of the global distortion

Similarly to the analysis of the global distortion performed in Section 4.2.1.3, the global rate-allocation problem consists in finding the sets $\{\lambda_j^*\}$ and $\{\Delta_j^*\}$ of optimal parameters which solve

$$\{\lambda_j^*\},\{\Delta_j^*\} = \begin{array}{ll} \arg\min & D\left(\{\lambda_j\},\{\Delta_j\}\right) \\ \text{subject to} & R\left(\{\lambda_j\},\{\Delta_j\}\right) \leq R_c, \\ & \lambda_j > 0, \forall j \in \{0,\ldots,J-1\} \\ & \Delta_j > 0, \forall j \in \{0,\ldots,J-1\} \end{array}. \tag{4.54}$$

Again, we need to express the mean global distortion D and the global coding rate R as a function of the sets of regularizing parameters $\{\lambda_j\}$ and quantizing steps $\{\Delta_j\}$ for the on-board imaging chain presented Fig. 4.12.

Proposition 4. *If $\sigma_{w_{z,j}}$ verifies hypothesis (4.51) for each $j \in \{0,\ldots,J-1\}$, then the mean global distortion D of the imaging chain displayed Fig. 4.12 writes*

$$D(\{\lambda_j\},\{\Delta_j\}) = \sum_{j=0}^{J-1} \frac{\pi_j a_j \lambda_j^2}{(1+\lambda_j)^2}\sigma^2_{w_{x,j}} + \frac{\pi_j a_j}{(1+\lambda_j)^2}\sigma^2_{w_{z,j}} + \pi_j a_j \frac{\Delta_j^2}{12}, \tag{4.55}$$

where

$$a_j = \frac{N_j}{N}, \tag{4.56}$$

is the weight of the subband j in the whole image.

Proof. As shown previously, the global distortion can be written as

$$D(\{\lambda_j\},\{\Delta_j\}) = \frac{1}{N}\sum_{j=0}^{J-1} \pi_j E\left(\|W_{x,j} - W_{\hat{x},j}\|^2\right), \tag{4.57}$$

where π_j are weighting coefficients which depend on the filters and the decimation factors used in the wavelet transform [Usevitch 1996]. In the case of the studied imaging chain displayed Fig. 4.12, the final image is the output of the coding/decoding and, from (4.48), writes

$$w_{\hat{x},j} = \frac{w_{x,j}}{1+\lambda_j} + w_{\varepsilon,j}. \tag{4.58}$$

From (4.57), (4.58) and using the moments decorrelation hypothesis (4.9), we deduce the global distortion

$$D(\{\lambda_j\},\{\Delta_j\}) = \frac{1}{N}\sum_{j=0}^{J-1} \frac{\pi_j\lambda_j^2}{(1+\lambda_j)^2} E\left(\|W_{x,j}\|^2\right) + \pi_j E\left(\|W_{\varepsilon,j}\|^2\right). \tag{4.59}$$

Finally, the global distortion (4.59) can be further developed using the results (4.53) to obtain the expression (4.55). □

The second part of the global rate-allocation problem (4.54) requires the expression of the global coding rate R. This rate can be expressed as the weighted sum of the rate in each subband R_j, estimated by its entropy [Shannon 1948]

$$R(\{\lambda_j\},\{\Delta_j\}) = \sum_{j=0}^{J-1} a_j R_j(\Delta_j), \tag{4.60}$$

where a_j is given in (4.56) and

$$R_j(\Delta_j) = -\sum_{m=-\infty}^{+\infty} P_{w_{\tilde{x},j}}(m,\Delta_j)\log_2\left(P_{w_{\tilde{x},j}}(m,\Delta_j)\right), \tag{4.61}$$

where $P_{w_{\tilde{x},j}}(m,\Delta_j)$ is the probability to get the symbol m which depends on the density probability function $p_{w_{\tilde{x},j}}$ of the subband $w_{\tilde{x},j}$ and on the quantizing step Δ_j

$$P_{w_{\tilde{x},j}}(m,\Delta_j) = \int_{m\Delta_j-\frac{\Delta_j}{2}}^{m\Delta_j+\frac{\Delta_j}{2}} p_{w_{\tilde{x},j}}(w_{\tilde{x},j})dw_{\tilde{x},j}. \tag{4.62}$$

From Section 4.1.1, we assume that each wavelet subband follows the generalized centered Gaussian distribution law defined in (4.1), where the parameters $\sigma^2_{w_{\tilde{x},j}}$ and $\alpha_{w_{\tilde{x},j}}$ of the distribution law will be estimated using the kurtosis-based technique proposed in [Kasner 1999].

Proposition 5. *The global rate-distortion optimization problem (4.54) can be solved by maximizing*

$$L(\tau) = \begin{array}{ll} \inf & \phi_\tau(\{\Delta_j\},\{\lambda_j\}) \\ & \lambda_j > 0,\, \forall j \in \{0,\dots,J-1\} \\ & \Delta_j > 0,\, \forall j \in \{0,\dots,J-1\} \end{array}, \tag{4.63}$$

with respect to $\tau > 0$ and where

$$\phi_\tau(\{\Delta_j\},\{\lambda_j\}) = \sum_{j=0}^{J-1} \frac{\pi_j a_j \lambda_j^2}{(1+\lambda_j)^2}\sigma^2_{w_{x,j}} + \frac{\pi_j a_j}{(1+\lambda_j)^2}\sigma_z^2 + \pi_j a_j \frac{\Delta_j^2}{12} + \tau\left(\sum_{j=0}^{J-1} a_j R_j(\Delta_j) - R_c\right). \tag{4.64}$$

Proof. This proof is similar to the one given in proposition 2 where the global distortion D is now given by (4.55). □

We detail in the next part how to solve problem (4.54) for the on-board imaging chain.

4.2.2.4 Global rate-distortion-denoising optimization

Using proposition 5, the optimization problem (4.54) becomes

$$\{\Delta_j^*\},\{\lambda_j^*\} = \max_{\tau>0}\left(\begin{array}{ll} \inf & \phi_\tau(\{\Delta_j\},\{\lambda_j\}) \\ & \lambda_j > 0,\, \forall j \in \{0,\dots,J-1\} \\ & \Delta_j > 0,\, \forall j \in \{0,\dots,J-1\} \end{array}\right). \tag{4.65}$$

where ϕ_τ is given in (4.64). We can show that a solution of problem (4.65) exists and is unique (see Appendix A.3). To find this solution, we propose to use the technique presented in Section 4.2.2.4 and based on the resolution of the simultaneous equations obtained from the KKT conditions [Kuhn 1951] of problem (4.65).

Proposition 6. *The KKT conditions of problem (4.65) admits only one solution $(\{\lambda_j^*\}, \{\Delta_j^*\}, \tau^*)$ which verifies*

$$\lambda_j^* = \frac{\sigma_z^2}{\sigma^2_{w_{x,j}}}, \quad \forall j \in \{0,\dots,J-1\} \tag{4.66}$$

$$\frac{\pi_j \Delta_j^*}{6} + \tau^* \frac{\partial R_j}{\partial \Delta_j}(\Delta_j^*) = 0, \quad \forall j \in \{0,\dots,J-1\} \tag{4.67}$$

$$\sum_{j=0}^{J-1} a_j R_j(\Delta_j^*) = R_c. \tag{4.68}$$

Proof. From the KKT conditions of problem (4.65), we get (see Appendix A.3)

$$\frac{\partial\phi(\Delta_j^*,\lambda_j^*,\tau^*)}{\partial\Delta_j} = \frac{a_j\pi_j\Delta_j^*}{6} + \tau^* a_j \frac{\partial R_j}{\partial\Delta_j}(\Delta_j^*) = 0 \tag{4.69}$$

$$\frac{\partial\phi(\Delta_j^*,\lambda_j^*,\tau^*)}{\partial\tau} = \sum_{j=0}^{J-1} a_j R_j(\Delta_j^*) - R_c = 0 \tag{4.70}$$

$$\frac{\partial\phi(\Delta_j^*,\lambda_j^*,\tau^*)}{\partial\lambda_j} = \frac{12a_j\pi_j\lambda_j^*\sigma^2_{w_{x,j}} - 12a_j\pi_j\sigma_z^2}{6(1+\lambda_j^*)^3} = 0 \tag{4.71}$$

$$\tag{4.72}$$

with

$$\frac{\partial R_j}{\partial\Delta_j}(\Delta_j) = -\frac{1}{\log(2)} \sum_{m=-\infty}^{+\infty} \left[1 + \log\left(P_{w_{\tilde{x},j}}(m,\Delta_j)\right)\right] \times \left[p_{w_{\tilde{x},j}}\left(m\Delta_j + \frac{\Delta_j}{2}\right)\left(m+\frac{1}{2}\right) - p_{w_{\tilde{x},j}}\left(m\Delta_j - \frac{\Delta_j}{2}\right)\left(m-\frac{1}{2}\right)\right]. \tag{4.73}$$

The expression (4.66) and conditions (4.67) and (4.68) on the optimal parameters directly follow from the optimality conditions (4.69). The existence and uniqueness of these parameters is detailled in Appendix A.3. □

As we can see from (4.67) and (4.68), the parameters $\{\Delta_j^*\}$ and τ^* still can not be computed in closed-form and will be estimated numerically using binary search algorithms of precision $\rho = 0.1$. The case of the low frequency subband ($j = J-1$) will be also processed differently to prevent excessive quantizing on these coefficients. We set

$$\Delta_{J-1}^* = 1, \tag{4.74}$$

$$\lambda_{J-1}^* = \frac{\sigma_z^2}{\sigma^2_{w_{x,J-1}}}. \tag{4.75}$$

Finally, the joint optimization procedure for solving problem (4.54) is given in the Algorithm 2. We do not include here the results of this algorithm as we have already shown in Section 4.2.1.5 that the proposed method was efficient to formulate an estimation of the global distortion for the on-ground imaging chain. Using an on-board restoration does not however affect the reliability of the proposed method, as shown in Section 4.2.2.3. Instead, we will show some results of this algorithm in the section dedicated to the comparison of the performances of the three chains (on-ground, on-board and hybrid that we present in the next part).

4.2.3 Optimization of the hybrid chain

4.2.3.1 Presentation of the imaging chain

As mentioned in the beginning of Section 4.2.2.1, it may be interesting to extend the on-board imaging chain by adding a supplementary denoising step, after coding,

Algorithm 2 Global rate-distortion-denoising joint optimization algorithm for the on-board imaging chain

Set $\tau = 1$.
Set $\rho = 0.1$.
while $\left|\sum_{j=0}^{J-1} a_j R_j - R_c\right| > \rho$ **do**
  for j from 0 to $J-2$ **do**
    Set $\Delta_j = 1$.
    Compute the value of the regularizing parameter λ_j from (4.66).
    while $\left|\frac{\pi_j \Delta_j}{6} + \tau \frac{\partial R_j}{\partial \Delta_j}(\Delta_j)\right| > \rho$ **do**
      Increase the value of Δ_j.
      Compute the value of the regularizing parameter λ_j from (4.66).
    end while
  end for
  Set $\Delta_{J-1} = 1$.
  Compute the regularizing paramater λ_{J-1} from (4.75).
  if $\left|\sum_{j=0}^{J-1} a_j R_j - R_c\right| > \rho$ **then**
    Increase the value of τ.
  end if
end while
Output the optimal regularizing parameters $\{\lambda_j^*\}$.
Output the optimal quantizing steps $\{\Delta_j^*\}$.

to reduce the quantizing noise. This "hybrid" chain is depicted Fig. 4.13. The instrumental image y is still a deteriorated version of the original image x where an additive instrumental noise z has been added.

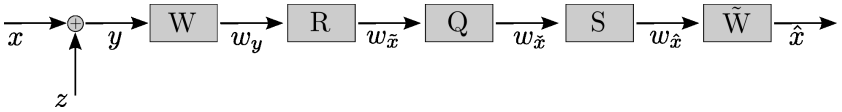


Figure 4.13: Considered hybrid imaging chain

The wavelet subbands of the instrumental image are again denoted $w_{y,j}$, $j \in \{0, \dots, J-1\}$ and their denoised version $w_{\tilde{x},j}$. The quantized version of these denoised subbands are denoted $w_{\check{x},j}$. An additional denoising algorithm **S** has been added at the end of the chain to reduce the coding noise. This Algorithm is similar to the one used for the operator **R** and writes

$$w_{\hat{x},j} = \frac{w_{\check{x},j}}{1+\mu_j}. \tag{4.76}$$

where $w_{\hat{x},j}$ is the final denoised subband and $\mu_j > 0$ is a regularizing parameter. Let $w_{b,j}$ be the coding error of the subband j

$$w_{b,j} = \mathbf{Q}(w_{\tilde{x},j}) - w_{\tilde{x},j}. \tag{4.77}$$

We have

$$\begin{aligned} w_{\check{x},j} = \mathbf{Q}(w_{\tilde{x},j}) &= w_{\tilde{x},j} + w_{b,j} \\ &= \frac{w_{y,j}}{1+\lambda_j} + w_{b,j} \\ &= \frac{w_{x,j}}{1+\lambda_j} + \frac{w_{z,j}}{1+\lambda_j} + w_{b,j}, \end{aligned}$$

and, from (4.76)

$$\begin{aligned} w_{\hat{x},j} &= \frac{w_{\check{x},j}}{1+\mu_j}, \\ &= \frac{w_{x,j}}{(1+\lambda_j)(1+\mu_j)} + \frac{w_{z,j}}{(1+\lambda_j)(1+\mu_j)} + \frac{w_{b,j}}{1+\mu_j}, \\ &= \frac{w_{x,j}}{(1+\lambda_j)(1+\mu_j)} + w_{\varepsilon,j}. \end{aligned} \tag{4.78}$$

where $w_{\varepsilon,j} = \frac{w_{z,j}}{(1+\lambda_j)(1+\mu_j)} + \frac{w_{b,j}}{1+\mu_j}$ is the global error. We detail in the next part how to formulate an expression of the global distortion.

4.2.3.2 Decorrelation hypothesis

The decorrelation hypothesis (4.9) will also be used to compute the global distortion of the imaging chain presented Fig. 4.13. It is important to note that the hybrid chain is an extension of the on-board chain and only adds a post processing after

coding; all the on-board process remain therefore the same. From this remark, it seems clear that the condition for the validity of the decorrelation hypothesis remains identical and writes

$$\sigma_{w_{z,j}} > \frac{\Delta_j}{2}(1+\lambda_j). \tag{4.79}$$

If the instrumental noise z meets the dithering noise requirements, we have [Wannamaker 2000]

$$E\left[W_{\varepsilon,j}\right] = 0, \tag{4.80}$$

$$E\left[\|W_{\varepsilon,j}\|^2\right] = N_j \frac{\sigma^2_{w_{z,j}}}{(1+\lambda_j)^2(1+\mu_j)^2} + N_j \frac{\Delta_j^2}{12(1+\mu_j)^2}. \tag{4.81}$$

4.2.3.3 Analysis of the global distortion

The global rate-allocation problem consists now in finding the sets $\{\lambda_j^*\}$, $\{\mu_j^*\}$ and $\{\Delta_j^*\}$ of optimal parameters which solve

$$\{\lambda_j^*\},\{\mu_j^*\},\{\Delta_j^*\} = \begin{array}{rl} \arg\min & D\left(\{\lambda_j\},\{\mu_j\},\{\Delta_j\}\right) \\ \text{subject to} & R\left(\{\lambda_j\},\{\Delta_j\}\right) \leq R_c, \\ & \lambda_j > 0, \forall j \in \{0,\ldots,J-1\} \\ & \Delta_j > 0, \forall j \in \{0,\ldots,J-1\} \\ & \mu_j > 0, \forall j \in \{0,\ldots,J-1\} \end{array} . \tag{4.82}$$

In comparison to the analysis performed in Section 4.2.2.3, the expression of the global distortion D changes and is now function of two sets of regularizing parameters $\{\lambda_j\}$, $\{\mu_j\}$ and, of course, is also function of the quantizing steps $\{\Delta_j\}$. The expression of the global coding rate R remains however unchanged as the denoising step that we introduced acts after the coding step.

Proposition 7. *If $\sigma_{w_{z,j}}$ verifies hypothesis (4.79) for each $j \in \{0,\ldots,J-1\}$, then the mean global distortion D of the imaging chain displayed Fig. 4.13 writes*

$$D\left(\{\lambda_j\},\{\mu_j\},\{\Delta_j\}\right) = \sum_{j=0}^{J-1} \frac{\pi_j\left(\lambda_j+\mu_j+\lambda_j\mu_j\right)^2}{(1+\lambda_j)^2(1+\mu_j)^2}\sigma^2_{w_{x,j}} + \frac{\pi_j a_j}{(1+\lambda_j)^2(1+\mu_j)^2}\sigma^2_{w_{z,j}} + \pi_j a_j \frac{\Delta_j^2}{12(1+\mu_j)^2}, \tag{4.83}$$

where

$$a_j = \frac{N_j}{N}, \tag{4.84}$$

is the weight of the subband j in the whole image.

Proof. Using the orthogonality of wavelet subbands, the global distortion can be formulated as

$$D\left(\{\lambda_j\},\{\mu_j\},\{\Delta_j\}\right)=\frac{1}{N}\sum_{j=0}^{J-1}\pi_j E\left(\|W_{x,j}-W_{\hat{x},j}\|^2\right), \tag{4.85}$$

where π_j are weighting coefficients which depend on the filters and the decimation factors used in the wavelet transform [Usevitch 1996]. In the case of the studied imaging chain displayed Fig. 4.13, the final image is the output of the second denoising step which, from (4.78), writes

$$w_{\hat{x},j}=\frac{w_{x,j}}{(1+\lambda_j)(1+\mu_j)}+w_{\varepsilon,j}. \tag{4.86}$$

From (4.85), (4.86) and using the moments decorrelation hypothesis (4.9), we deduce the global distortion

$$D\left(\{\lambda_j\},\{\mu_j\},\{\Delta_j\}\right)=\frac{1}{N}\sum_{j=0}^{J-1}\frac{\pi_j\left(\lambda_j+\mu_j+\lambda_j\mu_j\right)}{(1+\lambda_j)^2(1+\mu_j)^2}E\left(\|W_{x,j}\|^2\right)+\pi_j E\left(\|W_{\varepsilon,j}\|^2\right). \tag{4.87}$$

Finally, the global distortion (4.87) can be further developed using the results (4.81) to obtain the expression (4.83). □

The second part of the global rate-allocation problem (4.54) requires the expression of the global coding rate R. As the on-board processes of the hybrid imaging chain remain the same, the coding rate R is given by (4.60) and (4.61).

Proposition 8. *The global rate-distortion optimization problem (4.82) can be solved by maximizing*

$$L(\tau)=\begin{array}{ll}\inf & \phi_\tau(\{\Delta_j\},\{\mu_j\},\{\lambda_j\}) \\ & \lambda_j>0,\ \forall j\in\{0,\dots,J-1\} \\ & \Delta_j>0,\ \forall j\in\{0,\dots,J-1\} \\ & \mu_j>0,\ \forall j\in\{0,\dots,J-1\}\end{array}, \tag{4.88}$$

with respect to $\tau>0$ and where

$$\begin{aligned}\phi_\tau(\{\Delta_j\},\{\mu_j\},\{\lambda_j\})=&\sum_{j=0}^{J-1}\frac{\pi_j a_j\left(\lambda_j+\mu_j+\lambda_j\mu_j\right)^2}{(1+\lambda_j)^2(1+\mu_j)^2}\sigma^2_{w_{x,j}}+\frac{\pi_j a_j}{(1+\lambda_j)^2(1+\mu_j)^2}\sigma^2_{w_{z,j}}\\&+\pi_j a_j\frac{\Delta_j^2}{12(1+\mu_j)^2}+\tau\left(\sum_{j=0}^{J-1}a_j R_j(\Delta_j)-R_c\right).\end{aligned} \tag{4.89}$$

Proof. This proof is similar to the one given in proposition 2 where the global distortion D is given by (4.83) and R is given by (4.60) and (4.61). □

We detail in the next part how to solve problem (4.82) for the hybrid imaging chain.

4.2.3.4 Global rate-distortion-denoising optimization

Using proposition 8, the optimization problem (4.82) becomes

$$\{\Delta_j^*\},\{\mu_j^*\},\{\lambda_j^*\} = \max_{\tau>0} \left(\begin{array}{ll} \inf & \phi_\tau(\{\Delta_j\},\{\mu_j\},\{\lambda_j\}) \\ & \lambda_j > 0,\, \forall j \in \{0,\dots,J-1\} \\ & \Delta_j > 0,\, \forall j \in \{0,\dots,J-1\} \\ & \mu_j > 0, \forall j \in \{0,\dots,J-1\} \end{array} \right). \tag{4.90}$$

where ϕ_τ is given in (4.89). The situtation here is slightly different than the on-ground or on-board chains since problem (4.82) does not have any solution (see Appendix A.4). This means that we are not able to optimize in the same time the parameters of the two restorations (on-board and on-ground) used by this chain. We chosed therefore to enforce the value of λ_j^* as the same than for the on-board chain and we deduce the conditions of the three other parameters (see Appendix A.4)

$$\lambda_j^* = \frac{\sigma_z^2}{\sigma_{w_{x,j}}^2}, \quad \forall j \in \{0,\dots,J-1\} \tag{4.91}$$

$$\mu_j^* = \frac{\Delta_j^{*2}}{12\sigma_{w_{x,j}}^2}\left(1+\frac{\sigma_z^2}{\sigma_{w_{x,j}}^2}\right), \quad \forall j \in \{0,\dots,J-1\} \tag{4.92}$$

$$\frac{\pi_j \Delta_j^*}{6} + \tau^* \frac{\partial R_j}{\partial \Delta_j}(\Delta_j^*) = 0, \quad \forall j \in \{0,\dots,J-1\} \tag{4.93}$$

$$\sum_{j=0}^{J-1} a_j R_j(\Delta_j^*) = R_c. \tag{4.94}$$

As we can see from (4.93) and (4.94), the parameters $\{\Delta_j^*\}$ and τ^* still can not be computed in closed-form and will be estimated numerically using binary search algorithms of precision $\rho = 0.1$. The case of the low frequency subband ($j = J-1$) will be also processed differently to prevent excessive quantizing on these coefficients. We set

$$\Delta_{J-1}^* = 1, \tag{4.95}$$

$$\lambda_{J-1}^* = \frac{\sigma_z^2}{\sigma_{w_{x,J-1}}^2}, \tag{4.96}$$

$$\mu_{J-1}^* = \frac{1}{12\sigma_{w_{x,J-1}}^2}\left(1+\frac{\sigma_z^2}{\sigma_{w_{x,J-1}}^2}\right). \tag{4.97}$$

Since the on-board denoising parameter has been fixed, the optimization algorithm can be deduced from the one presented for the on-ground chain. We therefore get the suboptimal algorithm presented in Algorithm 3. The results of this algorithm are given in Section 4.3 which is dedicated to the comparison of the three imaging chains described in Section 4.2.1, 4.2.2 and 4.2.3.

Algorithm 3 Rate-distortion-denoising optimization algorithm for the hybrid imaging chain

Set $\tau = 1$.
Set $\rho = 0.1$.
while $\left|\sum_{j=0}^{J-1} a_j R_j - R_c\right| > \rho$ **do**
  for j from 0 to $J-2$ **do**
    Set $\Delta_j = 1$.
    Compute the value of the regularizing parameters λ_j from (4.91) and μ_j from (4.92).
    while $\left|\frac{\pi_j \Delta_j}{6} + \tau \frac{\partial R_j}{\partial \Delta_j}(\Delta_j)\right| > \rho$ **do**
      Increase the value of Δ_j.
      Compute the value of the regularizing parameters λ_j from (4.91) and μ_j from (4.92).
    end while
  end for
  Compute the quantizing step Δ_{J-1} from (4.95).
  Compute the regularizing paramaters λ_{J-1} from (4.96) and μ_{J-1} from (4.97).
  if $\left|\sum_{j=0}^{J-1} a_j R_j - R_c\right| > \rho$ **then**
    Increase the value of τ.
  end if
end while
Output the regularizing parameters $\{\lambda_j^*\}$ and $\{\mu_j^*\}$.
Output the quantizing steps $\{\Delta_j^*\}$.

4.3 Comparison of the three imaging chains

This part is dedicated to the comparison of the three chain (on-ground, on-board and hyrid) visually and in a rate-distortion sense. For this comparison, the reference image (displayed Fig. 3.1) has been noised with an additive white Gaussian noise whose standard deviation is equal to 10. The other parameters are the same than the ones described in Section 4.2.1.5. For each target rate, we simulate each imaging chain with the usual disjoint optimization technique in comparison to the proposed joint optimization algorithm.

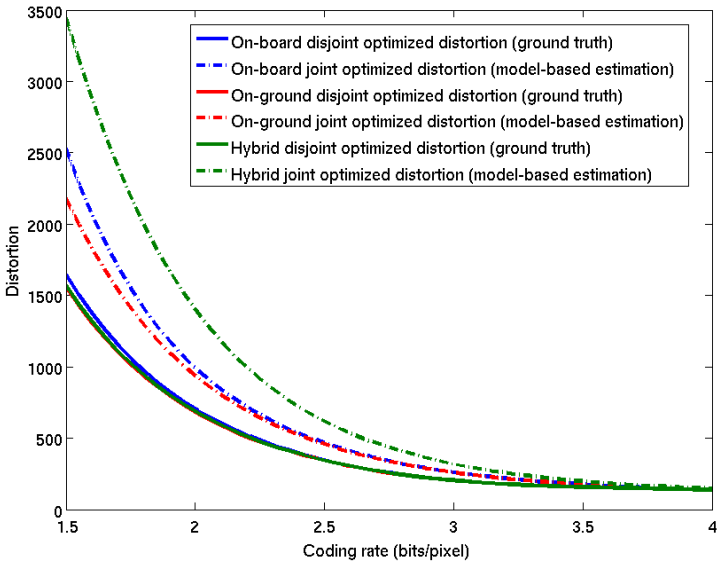


Figure 4.14: Comparison of the disjoint optimized distortion (ground truth) to joint optimized distortion (model-based estimation) for the three imaging chains, $\sigma_z = 10$.

The obtained rate-distortion curve is given Fig. 4.14. Since we simulate the case $\sigma_z = 10$, it is not suprising to observe that the joint optimization is slightly better than the disjoint optimization technique, in terms of global distortion, only for very high coding rates. This behavior is quite expected for the simulated level of instrumental noise since, as mentioned previously, the validity of the proposed method depends on the power of the instrumental noise. For $\sigma_z = 10$, we clearly know that the proposed approach will be valid only for high coding rates. However, for these rates, the coding step is almost transparent and therefore disjoint and joint optimized techniques are almost the same. At low coding rates, the decorrelation hypothesis does not hold anymore and the proposed method does not give a good estimation of the global distortion.

We also see on Fig. 4.14 that the on-board and on-ground chains give similar results, the on-ground chain being slightly better in term of global distortion. This is actually not suprising if we look at the estimation of the global distortion (4.16)

and (4.55), we see that one term is not attenuated by the regularizing term for the on-board chain. This remark actually leaded us to propose the hybrid chain. But we see on Fig. 4.14 that the joint optimization of this chain does not give satisfying results which is not suprising since the global optimization of this chain is not achievable (see Section 4.2.3.4).

The reconstructed images are given Fig. 4.15 to 4.18. The value of the peak signal-to-noise ration (PSNR) is given for indication. It is defines, for 12 bits dynamic images, by

$$PSNR(x, \hat{x}) = 20 \log_{10} \left(\frac{4095}{\frac{1}{N}\|x - \hat{x}\|_2} \right), \tag{4.98}$$

where N is the number of pixels, x is the reference image and $\hat{x}$ is the reconstructed final image.

The visual results are also similar, although we can observe on Fig. 4.16 a significant difference on the reconstructed images. On this zone, we observe that the on-board chain gives an image with less blur and artifacts than the ones obtained with the other chains. This result may however differ for other restoration algorithms since we used a Wiener like technique which is well adapted to process Gaussian noise but not coding noise.

We finally see that, visually, the reconstructed image with the joint optimization is better, for each chain, than the one reconstructed with the disjoint optimization technique. This result is actually quite surprising since the simulated coding rate is 2.5 bpp for which the dithering hypothesis does not hold anymore. This result is interesting and suggests that, even for medium coding rates, the correlation between the global error and the source may be negliged, such that our estimation of the global distortion also holds for this range of coding rates.

To conclude, we see that the obtained results point out once again that optimizing coding and denoising separately is suboptimal and that the problem of imaging chain design need to be treated in its globality. The proposed approach is interesting in this sense and allows to perform the optimization of the global chain, i.e. from the true scene to the final reconstructed image. Some works need however to be done to improve the proposed method and we address in the next section the question of extending the proposed approach to the current imaging chain used by the CNES.

4.4 Extension of the proposed method to the CNES imaging chain

The current imaging chain used by the CNES differs from the one we used in this chapter mainly on three points:

- the presence of the PSF which requires a deconvolution,
- the presence of the dead-zone on the quantizer,

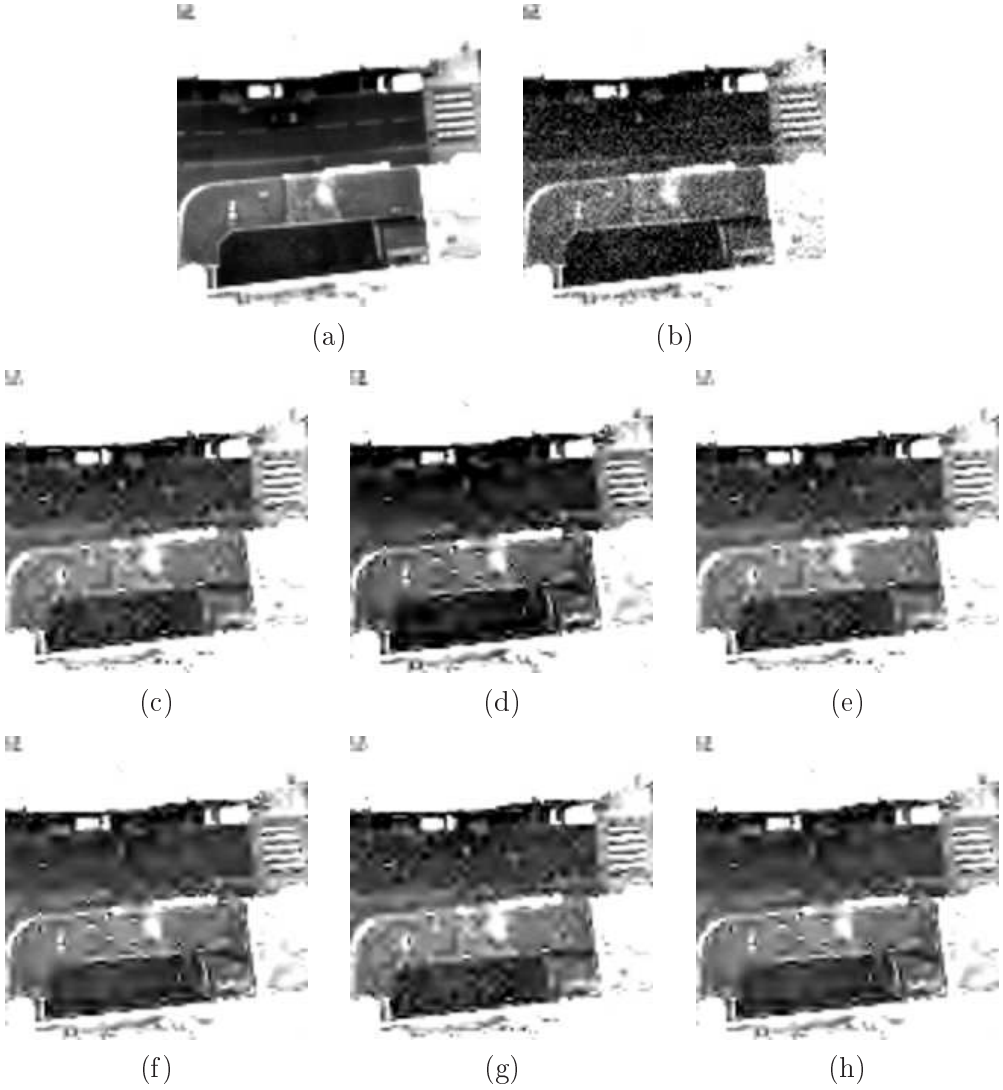


Figure 4.15: Visual comparison of reconstruction results. Displayed images have a size of 200×200 pixels. (a) is the reference image, (b) is the noisy observed image ($PSNR = 52.25\ dB$). (c) and (d) are the images reconstructed with the parameters obtained respectively by the disjoint minimization of the ground truth distortion ($PSNR = 46.95\ dB$) and by the joint optimization of the estimated distortion ($PSNR = 45.81\ dB$) for the *on-board* chain. (e) and (f) are the images reconstructed with the parameters obtained respectively by the disjoint minimization of the ground truth distortion ($PSNR = 46.99\ dB$) and by the joint optimization of the estimated distortion ($PSNR = 44.47\ dB$) for the *hybrid* chain. (g) and (h) are the images reconstructed with the parameters obtained respectively by the disjoint minimization of the ground truth distortion ($PSNR = 47.01\ dB$) and by the joint optimization of the estimated distortion ($PSNR = 45.76\ dB$) for the *on-ground* chain. The coding rate is 2.5 bits/pixel. The image range has been extended to point up the image reconstruction artifacts.

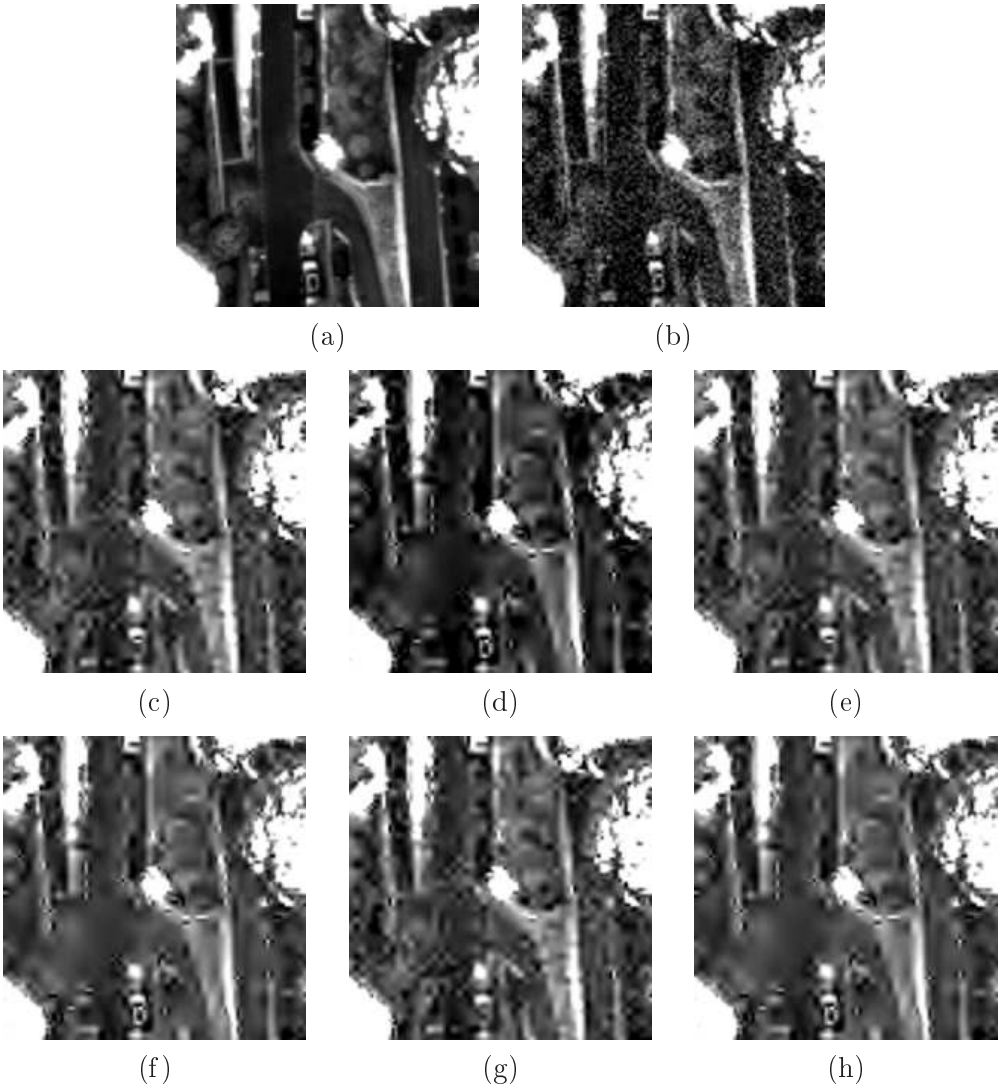


Figure 4.16: Visual comparison of reconstruction results. Displayed images have a size of 200×200 pixels. (a) is the reference image, (b) is the noisy observed image ($PSNR = 52.25\ dB$). (c) and (d) are the images reconstructed with the parameters obtained respectively by the disjoint minimization of the ground truth distortion ($PSNR = 46.95\ dB$) and by the joint optimization of the estimated distortion ($PSNR = 45.81\ dB$) for the *on-board* chain. (e) and (f) are the images reconstructed with the parameters obtained respectively by the disjoint minimization of the ground truth distortion ($PSNR = 46.99\ dB$) and by the joint optimization of the estimated distortion ($PSNR = 44.47\ dB$) for the *hybrid* chain. (g) and (h) are the images reconstructed with the parameters obtained respectively by the disjoint minimization of the ground truth distortion ($PSNR = 47.01\ dB$) and by the joint optimization of the estimated distortion ($PSNR = 45.76\ dB$) for the *on-ground* chain. The coding rate is 2.5 bits/pixel. The image range has been extended to point up the image reconstruction artifacts.

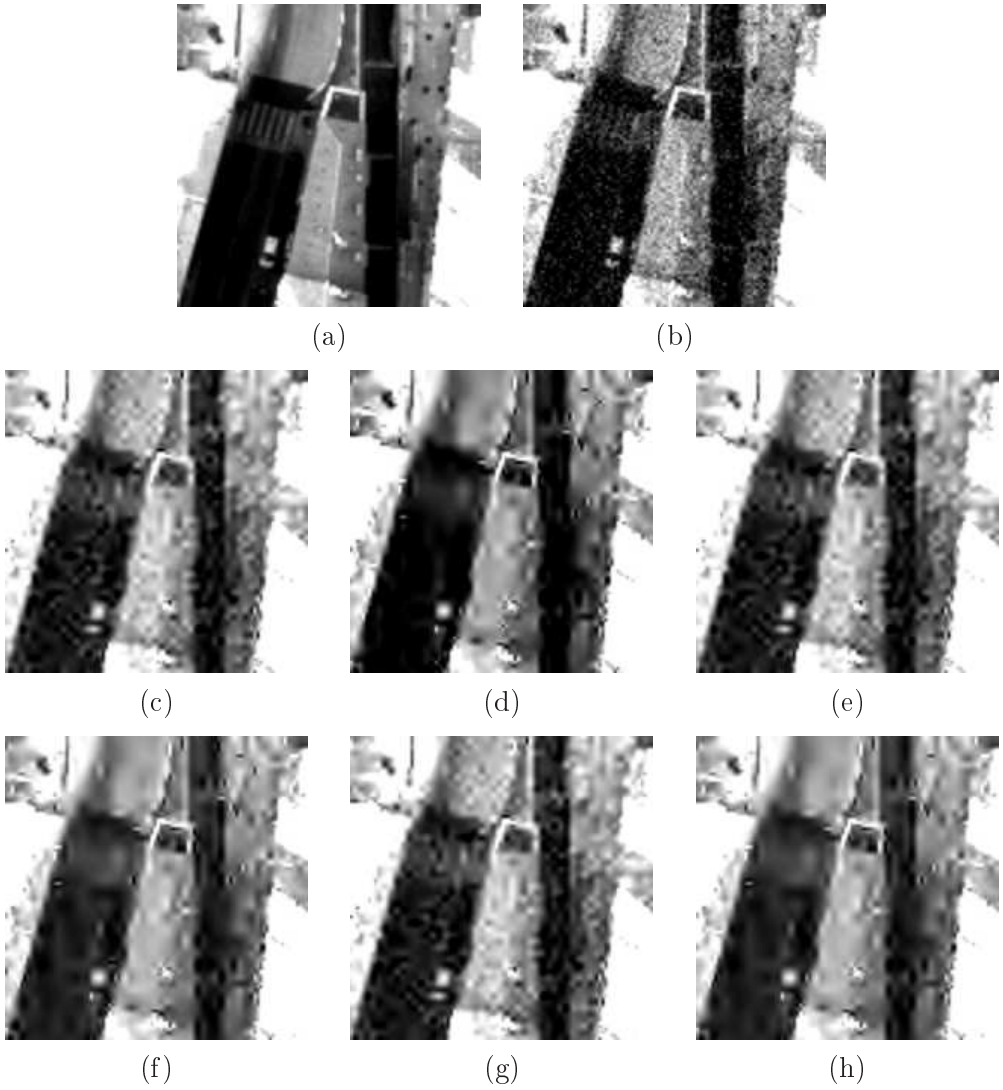


Figure 4.17: Visual comparison of reconstruction results. Displayed images have a size of 200×200 pixels. (a) is the reference image, (b) is the noisy observed image ($PSNR = 52.25\ dB$). (c) and (d) are the images reconstructed with the parameters obtained respectively by the disjoint minimization of the ground truth distortion ($PSNR = 46.95\ dB$) and by the joint optimization of the estimated distortion ($PSNR = 45.81\ dB$) for the *on-board* chain. (e) and (f) are the images reconstructed with the parameters obtained respectively by the disjoint minimization of the ground truth distortion ($PSNR = 46.99\ dB$) and by the joint optimization of the estimated distortion ($PSNR = 44.47\ dB$) for the *hybrid* chain. (g) and (h) are the images reconstructed with the parameters obtained respectively by the disjoint minimization of the ground truth distortion ($PSNR = 47.01\ dB$) and by the joint optimization of the estimated distortion ($PSNR = 45.76\ dB$) for the *on-ground* chain. The coding rate is 2.5 bits/pixel. The image range has been extended to point up the image reconstruction artifacts.

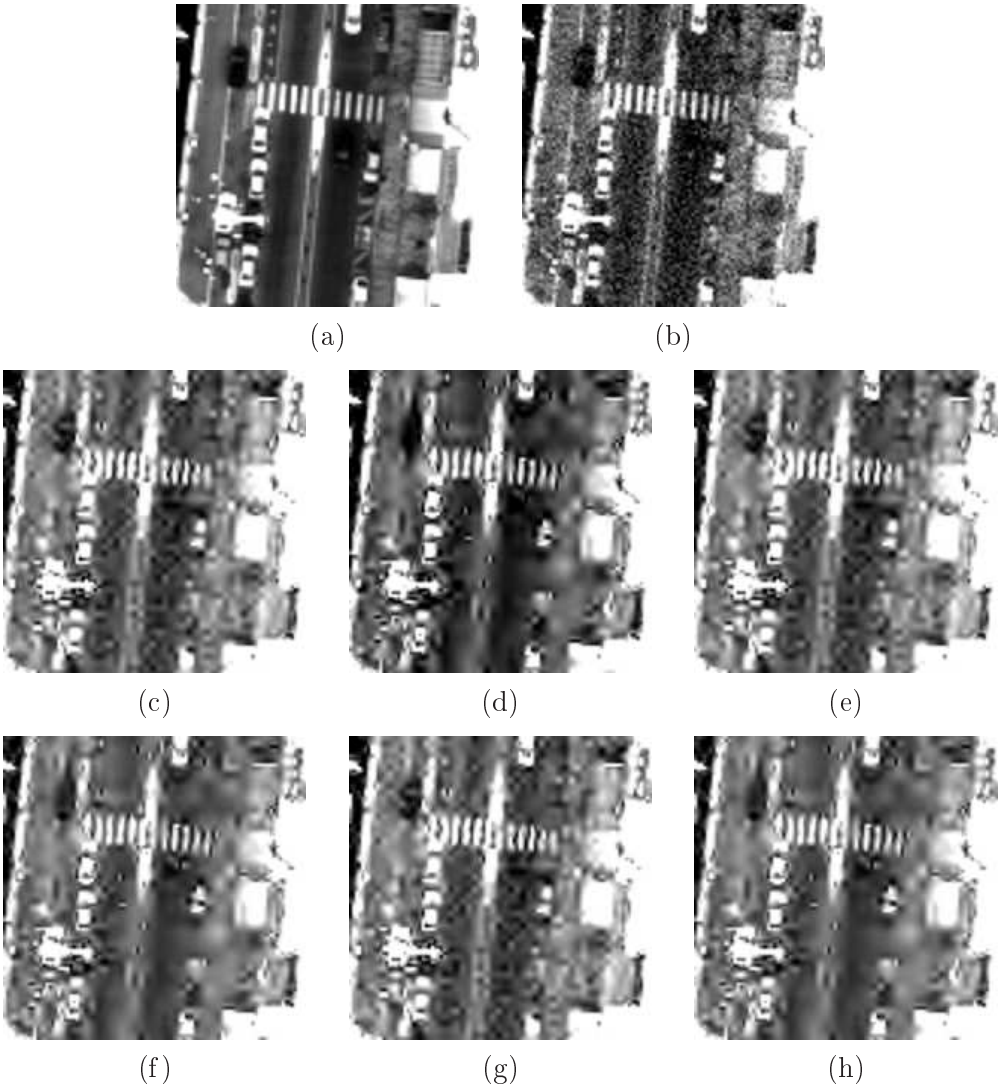


Figure 4.18: Visual comparison of reconstruction results. Displayed images have a size of 200×200 pixels. (a) is the reference image, (b) is the noisy observed image ($PSNR = 52.25\ dB$). (c) and (d) are the images reconstructed with the parameters obtained respectively by the disjoint minimization of the ground truth distortion ($PSNR = 46.95\ dB$) and by the joint optimization of the estimated distortion ($PSNR = 45.81\ dB$) for the *on-board* chain. (e) and (f) are the images reconstructed with the parameters obtained respectively by the disjoint minimization of the ground truth distortion ($PSNR = 46.99\ dB$) and by the joint optimization of the estimated distortion ($PSNR = 44.47\ dB$) for the *hybrid* chain. (g) and (h) are the images reconstructed with the parameters obtained respectively by the disjoint minimization of the ground truth distortion ($PSNR = 47.01\ dB$) and by the joint optimization of the estimated distortion ($PSNR = 45.76\ dB$) for the *on-ground* chain. The coding rate is 2.5 bits/pixel. The image range has been extended to point up the image reconstruction artifacts.

- and the denoising which is usually performed using shrinkage estimators instead of Wiener like techniques.

Introducing a deconvolution in the restoration algorithm that we used may be the easiest point to achieve. A deconvolution is usually performed in the Fourier domain and we expressed our global distortion in the wavelet domain. Moving from one domain to the other one may be difficult so one way to include this deconvolution is to use a wavelet packet denoising such that the variation of frequency inside a packet is low enough to be approximated by a constant. The deconvolution could then be approximated, for each packet, as a division by this constant.

The presence of a dead-zone in the quantizer is also a point that may be addressed. Theoretically, the dead-zone of the quantizer prevents the moments of the global error to be decorrelated to the moments of the source, as the dithering hypothesis requires an equally spaced quantizer. We are however confident that the correlation introduced by this dead-zone may be negliged such that the proposed approach can still be applied.

The main difficulty for extending this work to the imaging chain used by the CNES comes from the use of shinkage estimators. The non-linearity of these estimators makes our approach very difficult to extend to this case. Moreover, the lack of statistics on the reconstructed image of these estimators complexify the problem of global distortion estimation.

For these reasons, we propose in the next chapter a different approach to perform the global optimization of the chain.

4.5 Conclusions and perspectives

We studied in this chapter the global optimization of the chain from a theoretical point of view. We considered a simple case of imaging chain and we proposed a technique to estimate the global distortion. We also presented an algorithm to get the optimal coding and denoising parameters by minimizing the estimated global distortion with respect to the parameters of the chain, given a target coding rate.

We simulated this joint optimization technique on a satellite image and we showed this approach allows a significant improvement on the quality of the final image. In detail, our joint coding/denoising optimization approach can either allows to reach the same quality at lower rates or to improve the quality of the reconstructed final image for the same rates, in comparison to the image obtained using the classical disjoint optimization technique. The main conclusion obtained in this chapter is that the quality of the final image can be highly improved if we address the problem of the satellite imaging chain optimization in its globality and the proposed method is interesting in this sense.

We also developed our study to three configurations of the imaging chain where the restoration is either performed after coding, before coding or splitted in two parts: One part before coding and one part after coding. The comparison of these

three imaging chains showed that it is more interesting, in term of image quality, to place the restoration before coding, i.e. on-board of the satellite.

The imaging chain that we considered remains however simple and is far from the true satellite imaging chain which is much more complex. We discussed in Chapter 4.4 the main differences between the considered imaging chain and the system currently used by the CNES. The main difficulty to extend our method to that chain comes from the shrinkage-based restoration algorithm used by the CNES. Due to the lack of statistics on this type of algorithm, it seems highly difficult to formulate an expression of the final image. This however may be achieved if one allows to introduce more prior information that we used in this chapter.

Chapter 5

Numerical optimization of the chain

In the previous chapter we presented a method to perform, under simplifying hypotheses, a global joint optimization of the imaging chain which showed significant improvements on the visual quality of the final image. This method is however difficult to extend to the true imaging chain of a satellite, due to the non-stationarity of the instrumental noise, the non-linearity of the restoration technique and the presence of a dead-zone on the quantizer.

Although we are not able to express the global distortion as a function of the parameters of the chain, we will show in Section 5.1 that a global optimization can be approximately performed by simply shifting the position of the restoration in the chain. Tuning the parameters of the restoration is however theoretically difficult so we propose in this part to address this question numerically. This chapter focusses then on the global study of the satellite imaging chain, but mainly from a numerical point of view. We will first present in Section 5.1 numerical experiments to improve the quality of the final image by changing the position and the technique used for the restoration step. For visual considerations, we will show then in Section 5.2 how to deal with the structured artifacts of the coding noise. We conclude in Section 5.4 and give perspectives of the study.

5.1 Global optimization using on-board restoration

As mentioned in the introduction of the thesis (see Section 1.1), the initial global optimization problem consists in finding the optimal coding/decoding C^* and restoration T^* which minimizes on average some measure D of the distance between the true scene x and the restored final image $\hat{x} = T(C(y))$, under the constraint that the coding rate $R(C(y))$ does not exceed the target coding rate

$$C^*, T^* = \begin{array}{rl} \arg\min & E\left[D(x, T(C(y)))\right] \\ \text{subject to} & C, T \\ & R(C(y)) \leq R_c \end{array} . \tag{5.1}$$

Problem (5.1) is highly complex to solve as it looks for the optimal coding C^* and restoration T^* without any knowledge on the true image x and for any distance D. Clearly, solving (5.1) is very difficult to achieve in a general context. The authors of [Wolf 1970] have however shown that some simplifications can be made if the

distance D is the mean square error (MSE). The main result of [Wolf 1970] states that, in the case of the MSE, the global distortion can be separated in two terms as follows

$$D = E\left[\|x - T(C(y))\|_2^2\right] = E\left[\|x - E[x|y]\|_2^2\right] + E\left[\|E[x|y] - T(C(y))\|_2^2\right], \quad (5.2)$$

where $E[x|y]$ is the conditional expectation of the original image x knowing the noisy one y. The image $E[x|y]$ is the best (in the MSE sense) estimator of the original image x from y. As this image does not depend on the on-ground restoration or the compression technique used, the minimal distortion D^* then writes [Wolf 1970]

$$D^* = E\left[\|x - E[x|y]\|_2^2\right] + \begin{array}{ll} \min & E\left[\|E[x|y] - T(C(y))\|_2^2\right] \\ \text{subject to} & C, T \end{array}. \quad (5.3)$$

We see that the global distortion can be expressed and optimized with respect to the image $E[x|y]$ instead of the original image x. Note that the problem (5.3) is not simpler to solve as the computation of the image $E[x|y]$ is usually not accessible.

As mentioned previously, the image $E[x|y]$ represents the restoration of the true image x from the instrumental one y. It is then very tempting to think that this ideal image is actually the result of the restoration T, moved on-board of the satellite, i.e. before coding (see Fig. 5.1). From this remark, we then propose to consider the MSE as the distance D and to use the results of [Wolf 1970] on the problem (5.1). We further replace $E[x|y]$ by $T(y)$ such that the global optimization problem (5.1) can be approximatively written as

$$C^*, T^* = \begin{array}{ll} \arg\min & E\left[\|T(y) - C(T(y))\|_2^2\right] \\ \text{subject to} & C, T \\ & R(C(T(y))) \leq R_c \end{array}. \quad (5.4)$$

It is certain that the problem (5.4) is not strictly equal to the initial optimization problem (5.1). Problem (5.4) seems however easier to treat as each variable can almost be optimized separately. If T is fixed, problem (5.4) looks then for the optimal coder C^* which minimizes the coding error under the constraint that the coding rate does not exceed the target coding rate. This problem is well-known and referred as the coding rate-allocation problem [Shannon 1948] which has been addressed a lot in the coding community [Antonini 1992], [Ortega 1998], [Berger 1971] and references therein.

To be clear, the global joint optimization problem (5.1) is very difficult to address. But, in our opinion, we believe that moving the restoration on-board allows to optimize the global imaging chain by optimizing separately each process (restoration and coding) [1]. Moreover, the fact that each process needs to be optimized separately actually fits how these parts have been originally designed. This strengthens our

[1] If we go back to the theoretical study of the chain, in Section 4.2.2.4, we observe that the optimal parameters of the on-board chain are independent of each others, which is not the case of the on-ground chain

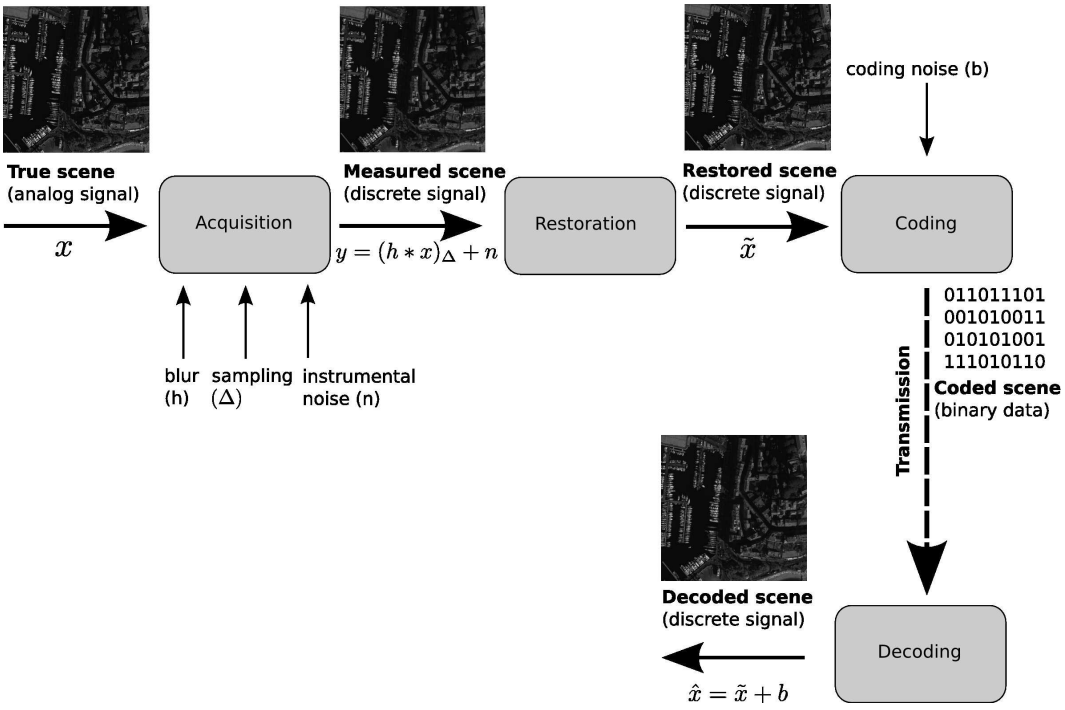


Figure 5.1: On-board restoration based satellite imaging chain.

idea that moving the restoration on-board is actually a reliable method to perform the global optimization. So one way (but again this is not the only one) to address the problem of global joint optimization (5.1) is to use an on-board restoration such that the global optimization problem can be approximatively splitted in two independent ones. The first problem is to optimize the on-board restoration such that it is close to $E[x|y]$. The second problem is to design a coder C which minimizes the coding error. As mentioned previously, the latter has been the focus of intense work in the imaging community. So the difficulty here is to evaluate how close to $E[x|y]$ is $T(y)$. As the ideal image $E[x|y]$ depends on the original image x and is therefore not accessible, we will simulate several state-of-the-art restoration algorithms and observe their impact on the global distortion and on the quality of the reconstructed image. This is the focus of the next part.

5.1.1 Comparison of on-board and on-ground chains

We are considering the on-board chain displayed Fig. 5.1 in comparison to the classical on-ground one illustrated Fig. 5.2 for several restoration algorithms.

For the simulation, the coding step is fixed and is performed using the method proposed in [CCSDS 2005] which is the basis of satellite embedded coding algorithms. For example, the technique implemented on-board of the recent PLEIADES-HR satellite is an extension of the method proposed in [CCSDS 2005]. To be con-

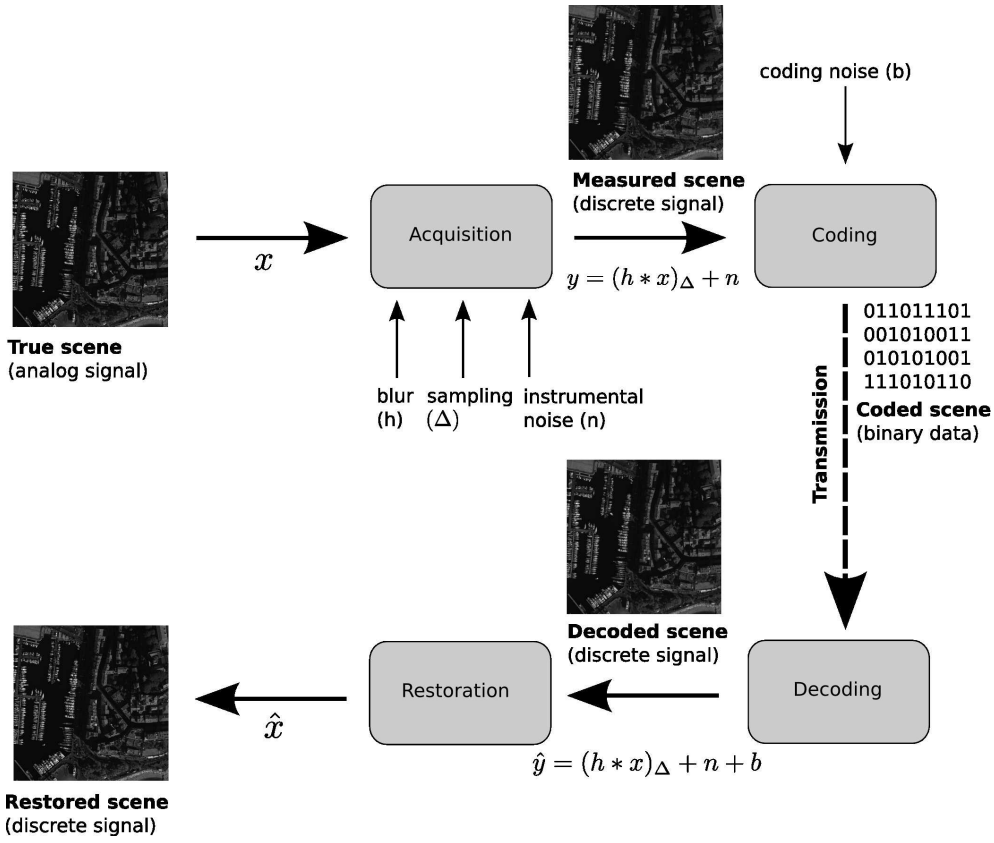


Figure 5.2: On-ground restoration based satellite imaging chain.

sistent with the technique used by the CNES, we only focus here on restoration techniques which process the image in two steps (we do not include the methods based on a variational framework such as [Bect 2004]) as follows. First, a direct deconvolution is performed using the target point spread function (PSF) provided by the CNES. This deconvolution tends to increase the power of the instrumental noise such that a post-processing denoising is always required as the second step. A wavelet packet decomposition [Kalifa 2003b] is usually used for this denoising as it fits the frequential characteristics of the deconvolved noise [Lier 2008]. However, another important point to take into account for an efficient denoising is the decrease rate of reconstruction error from the M largest wavelet coefficients [Patel 2009]. The faster the reconstruction error decreases, the better the denoising is. And on this point, a wavelet packet transform may not be optimal [Mallat 2008].

We propose here to perform the denoising using a variant of the wavelet transform named the Shearlet transform [Labate 2005]. A wavelet transform can be represented using a matrix with dyadic shifts and dilations as coefficients. As mentioned in Section 3.2.1, it is classically extented to the two dimensional case using separable wavelets which process each dimension of the image independently. The matrix representation of a two dimensional wavelet transform is therefore diagonal. The Shearlet transform presented in [Labate 2005] proposes instead to use a non-diagonal matrix and more specifically considers a "shear" matrix. A shear matrix is a matrix that combines operations along its rows and columns. This implies that a Shearlet transform uses combinations of shifts and dilations of each dimension of the image. This offers the ability to capture oriented details and is, among the contourlets [Do 2005] and the curvelets [Candès 2006a], an optimal transform (in term of reconstruction error decreasing rate with respect to the number of retained coefficients) for the representation of images [Patel 2009]. A deconvolution method based on the Shearlet transform has been proposed in [Patel 2009]. We will therefore compare the method [Patel 2009] to the current state-of-the-art restoration methods such as the ForWarRD method [Neelamani 2004], which performs a deconvolution followed by a regularization in both the Fourier and wavelet domains, or the method based on a Stein block thresholding [Chesneau 2010] which performs the regularization in the Vaguelet-Wavelet domain followed by an adaptive block thresholding.

We simulate both on-board and on-ground chains on the image presented Fig. 5.3 using the mentioned restoration algorithms. The reconstructed images will be compared to the ones provided by the CNES which, as mentioned in Section 3.3.2, uses an on-ground restoration based on a direct deconvolution followed by a wavelet packet thresholding. For the numerical experiments, the threshold parameters have been chosen such that the MSE is minimized. An exhaustive search of these parameters has been used to achieve this goal. In this simulation, the original image x is known and the MSE can thus be computed. Note that in a real environment, unbiased estimators of the MSE exist and do not require the knowledge of the true image [Ramani 2008]. Other estimators such as generalized cross validation (GCV) techniques [Golub 1979] may also be used.

Figure 5.3: Reference image, Cannes harbour (12 bits, 30 cm resolution, 1024×1024 pixels).

The quality of the reconstruction results will be estimated both visually and numerically using the PSNR criterion defined in (4.98). To evaluate visually the performances of these algorithms, we will only display the reconstructed images for the acquisition parameters described by the operating point 62 (whose SNR is 30-100 and target coding rate is 2.5 bpp) in Table 3.1, page 35. This operating point is very interesting to visually test the efficiency of the restoration algorithms since it gives the worst-case simulation parameters: An instrumental noise with a high standard deviation (low SNR $30 - 100$) and a low coding rate (2.5 bits/pixel).

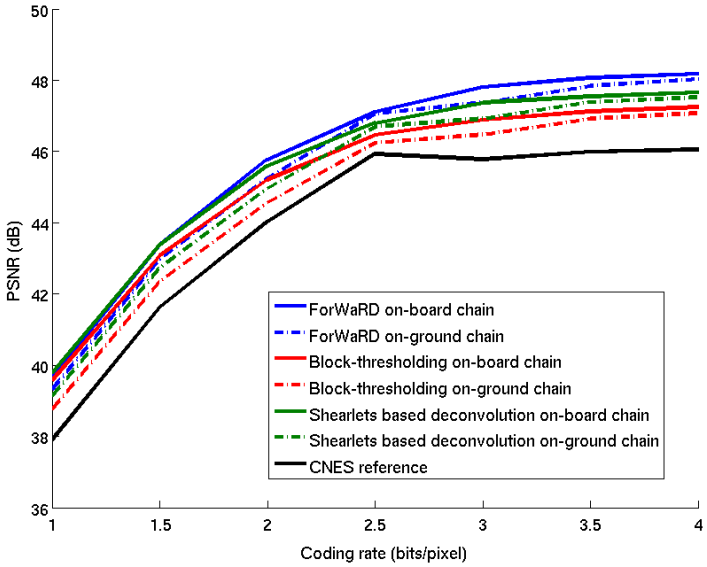


Figure 5.4: Rate-distortion comparison of on-board and on-ground chains in reference to the method currently used by the CNES. The simulated SNR is 30-100.

The comparison of the on-board and on-ground chains in a rate-distortion sense is given Fig. 5.4 to 5.6 for the different restoration algorithms and for different simulated signal-to-noise ratios. We can see that for the simulated restoration techniques, an on-board chain always performs better than its on-ground variants. At low coding rate, the difference between the two chains reaches almost 1 dB. We can also observe that each restoration technique outperforms the restoration technique used by the CNES in terms of PSNR. For a coding rate of 2.5 bpp, the improvement, in terms of PSNR, of these methods over the method of the CNES varies between 1 and 1.5 dB. Note that the PSNR of the method used by the CNES is almost constant after the coding rate of 2.5 bits/pixel as this technique leaves some residual noise to give the image a physical sense. This residual noise simulates the instrumental noise that one obtains at the output of a sensor. This phenomenon only appears from 2.5 bits/pixel, as at this rate the encoder starts to efficiently encode the instrumental noise instead of removing it. Also note that this image character-

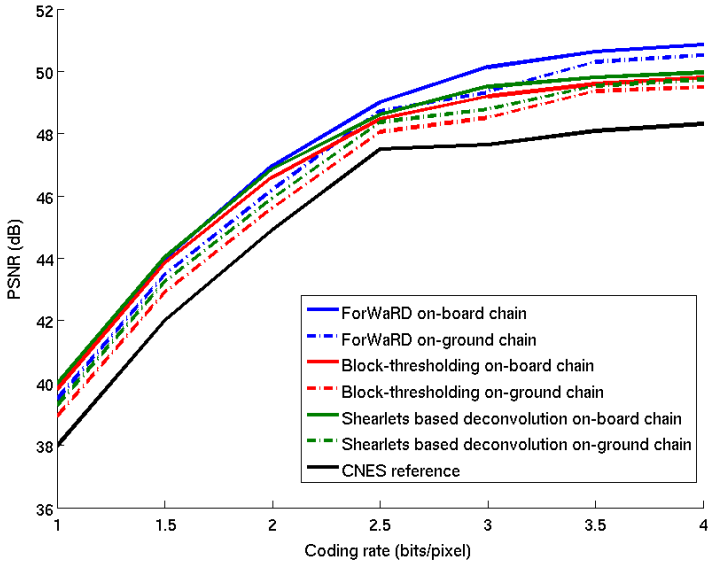


Figure 5.5: Rate-distortion comparison of on-board and on-ground chains in reference to the method currently used by the CNES. The simulated SNR is 30-150.

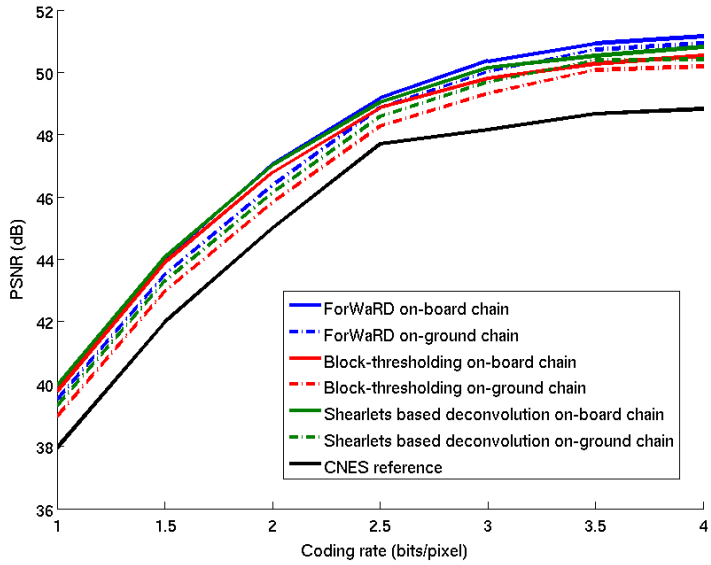


Figure 5.6: Rate-distortion comparison of on-board and on-ground chains in reference to the method currently used by the CNES. The simulated SNR is 50-150.

istic is highly appreciated by image analysis experts. This feature will be the basis of the method proposed in Section 5.3 to remove the coding artifacts inherent in wavelet-based compression systems.

Among the simulated techniques, the ForWarRD restoration algorithm [Neelamani 2004] gives the best PSNR for all coding rates. The difference with other methods is however very small such that it is difficult to conclude only from the rate-distorsion curves. To better evaluate the differences between these algorithms, we show visual results on the Fig. 5.7 to 5.10.

We can check on Fig. 5.7 for example that the on-board chain gives edges which are slightly more blurred than the on-ground chain (particularly visible around the edges of buildings). This is due to the fact that the edges of the image have been enhanced by the deconvolution. The high frequency subbands require then more bits to be properly encoded.

It is actually difficult to conclude on the difference between the two chains as they both give similar results, although the on-ground one seems to perform better on low intensity areas. For example, on Fig. 5.9, we see that the on-board chain reconstructs an image which is more blurred (see the small square element at the bottom of the figure) than the one we would have obtained with an on-ground chain (see also figure 5.7). The on-board chain presents however the advantage to separate the process of coding noise removal and we will exploit this ability later in Section 5.3.

Visually, the Stein block thresholding restoration technique [Chesneau 2010] does not give satisfying results and tends to oversmooth the image. If we observe the reconstructed images (Fig. 5.7 and 5.9 for example), we can verify that all the small details are lost. The ForWaRD method [Neelamani 2004] seems also to suffer from the same behavior and provides slightly smooth reconstructed images. The method based on the Shearlets [Patel 2009] seems to be slightly superior in term of image quality. This method give satisfying results and recover the small details of the image without giving too many artifacts. A deeper evaluation of the reconstructed images, by image analysis experts, may be however required to confirm this result.

Finally, we see that many coding artifacts still appear in the reconstructed images. This phenomenon is particularly visible on the reconstruction results of the on-board chain as the coding noise is not treated at all by this chain. The on-board chain may be therefore penalized by the presence of these artifacts, so we present in the next part some of the state-of-the-art processing methods to reduce these coding artifacts.

5.2 Coding noise removal

As mentioned in Section 5.1.1, the coding step of the imaging chain degrades the quality of the transmitted image by introducing structured artifacts. These artifacts are due to the quantizing process of the coder which sets to zero the wavelet coefficients of low magnitude. This action of quantizing to zero can be interpretated

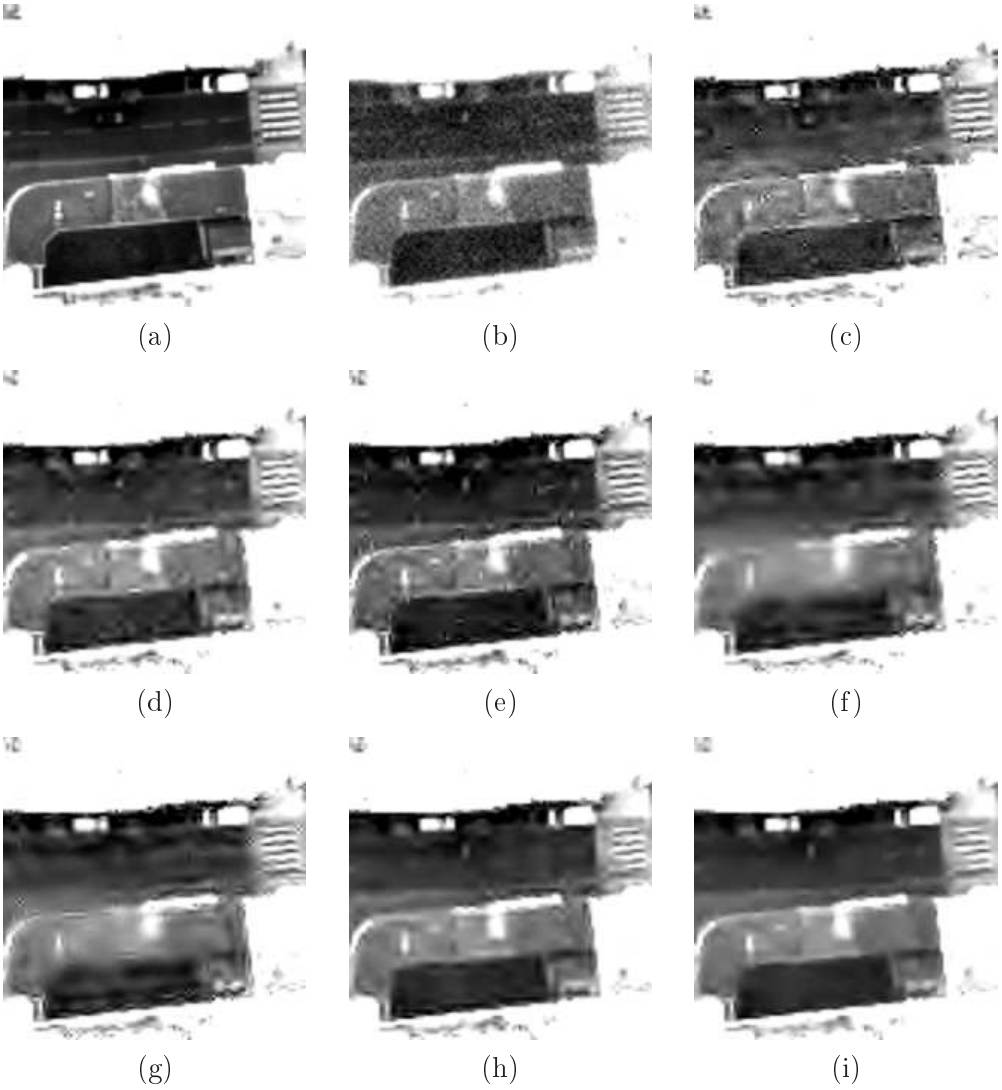


Figure 5.7: Visual comparison of on-board and on-ground chains. Displayed images have a size of 200×200 pixels. (a) is the reference image, (b) is the instrumental image (output of the acquisition, $PSNR = 32.69\ dB$), (c) is the reconstructed image provided by the CNES ($PSNR = 45.93\ dB$), (d) and (e) are the reconstructed images respectively from the Shearlets based on-board ($PSNR = 46.80\ dB$) and on-ground ($PSNR = 46.69\ dB$) chains, (f) and (g) are the reconstructed images respectively from the block thresholding based on-board ($PSNR = 46.46\ dB$) and on-ground ($PSNR = 46.24\ dB$) chains, (h) and (i) are the reconstructed images respectively from the ForWarRD based on-board ($PSNR = 47.11\ dB$) and on-ground ($PSNR = 47.05\ dB$) chains. The target rate is 2.5 bits/pixel and the simulated SNR is 30-100. The image range has been extended to point up the image reconstruction artifacts.

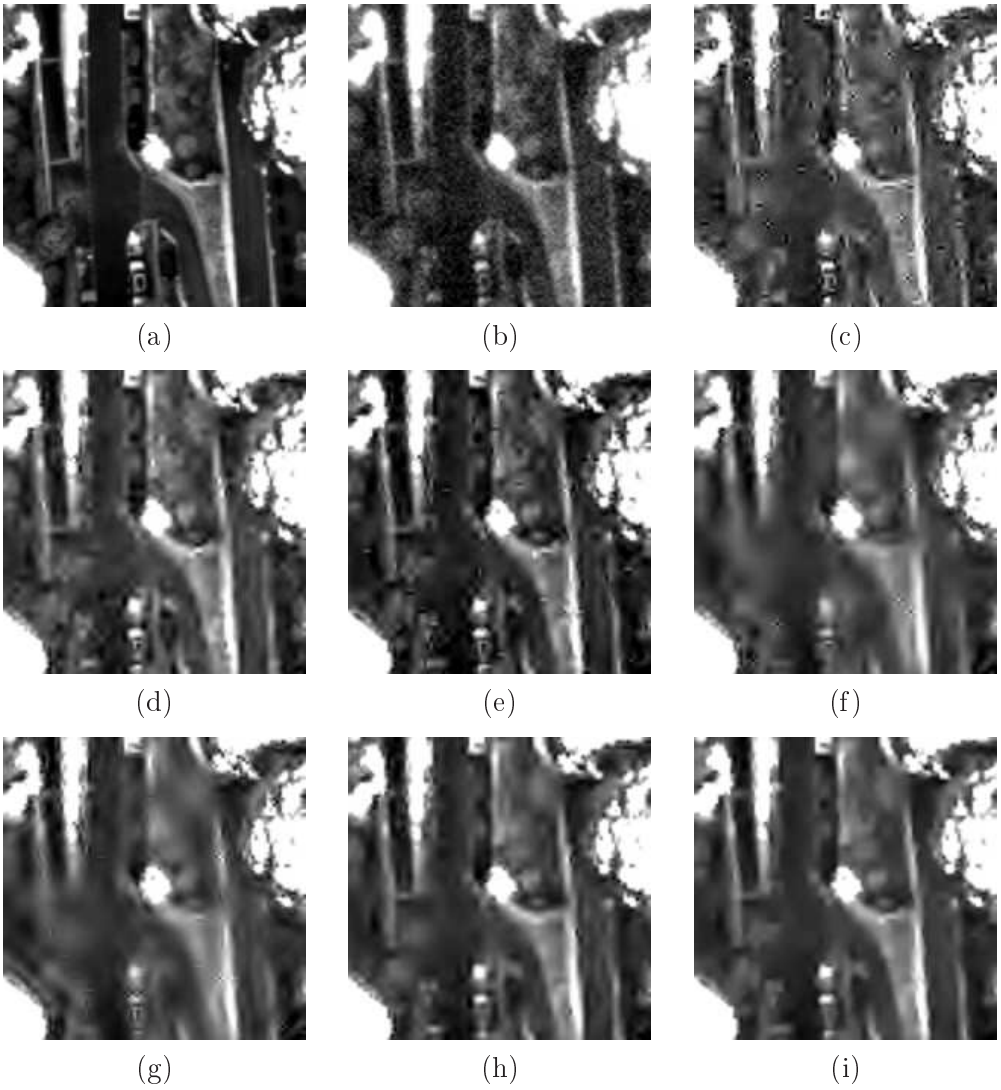


Figure 5.8: Visual comparison of on-board and on-ground chains. Displayed images have a size of 200×200 pixels. (a) is the reference image, (b) is the instrumental image (output of the acquisition, $PSNR = 32.69\ dB$), (c) is the reconstructed image provided by the CNES ($PSNR = 45.93\ dB$), (d) and (e) are the reconstructed images respectively from the Shearlets based on-board ($PSNR = 46.80\ dB$) and on-ground ($PSNR = 46.69\ dB$) chains, (f) and (g) are the reconstructed images respectively from the block thresholding based on-board ($PSNR = 46.46\ dB$) and on-ground ($PSNR = 46.24\ dB$) chains, (h) and (i) are the reconstructed images respectively from the ForWarRD based on-board ($PSNR = 47.11\ dB$) and on-ground ($PSNR = 47.05\ dB$) chains. The target rate is 2.5 bits/pixel and the simulated SNR is 30-100. The image range has been extended to point up the image reconstruction artifacts.

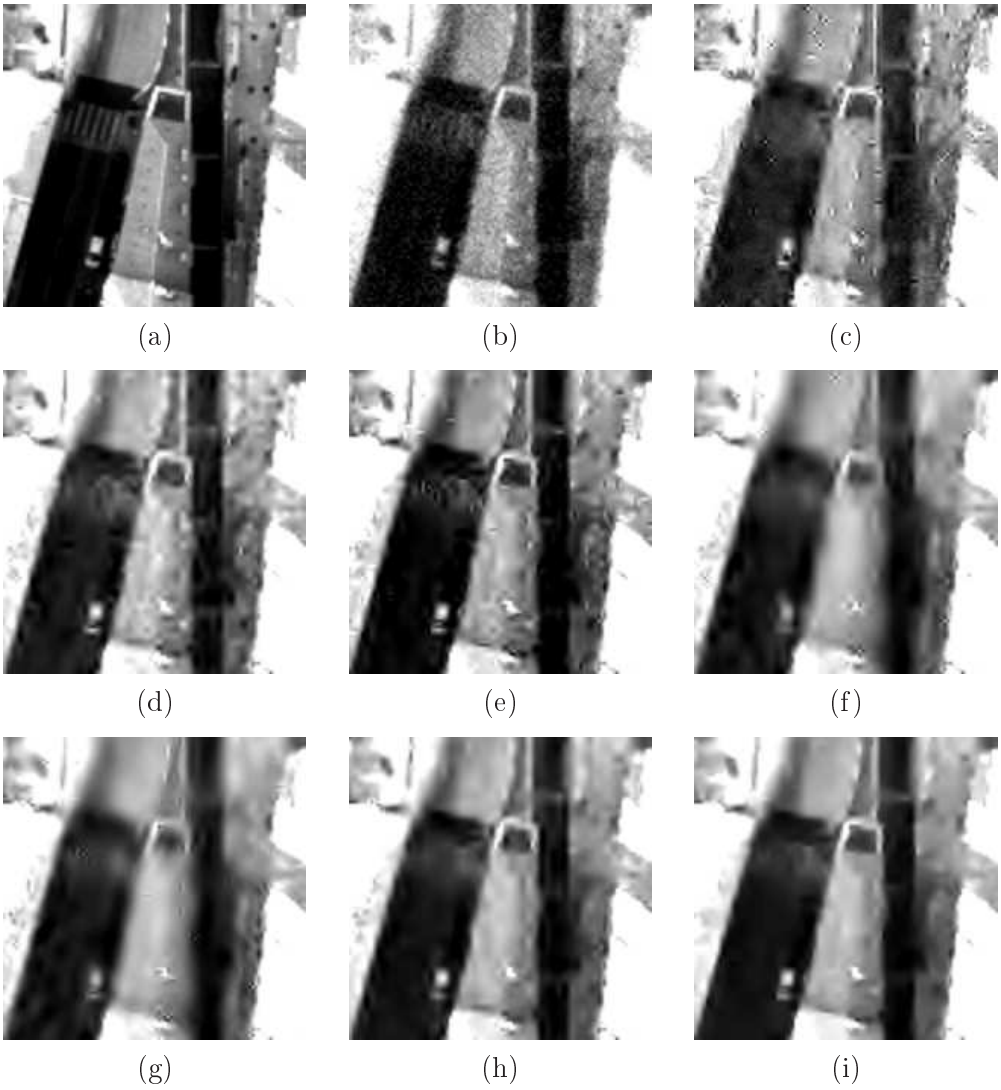


Figure 5.9: Visual comparison of on-board and on-ground chains. Displayed images have a size of 200×200 pixels. (a) is the reference image, (b) is the instrumental image (output of the acquisition, $PSNR = 32.69\ dB$), (c) is the reconstructed image provided by the CNES ($PSNR = 45.93\ dB$), (d) and (e) are the reconstructed images respectively from the Shearlets based on-board ($PSNR = 46.80\ dB$) and on-ground ($PSNR = 46.69\ dB$) chains, (f) and (g) are the reconstructed images respectively from the block thresholding based on-board ($PSNR = 46.46\ dB$) and on-ground ($PSNR = 46.24\ dB$) chains, (h) and (i) are the reconstructed images respectively from the ForWarRD based on-board ($PSNR = 47.11\ dB$) and on-ground ($PSNR = 47.05\ dB$) chains. The target rate is 2.5 bits/pixel and the simulated SNR is 30-100. The image range has been extended to point up the image reconstruction artifacts.

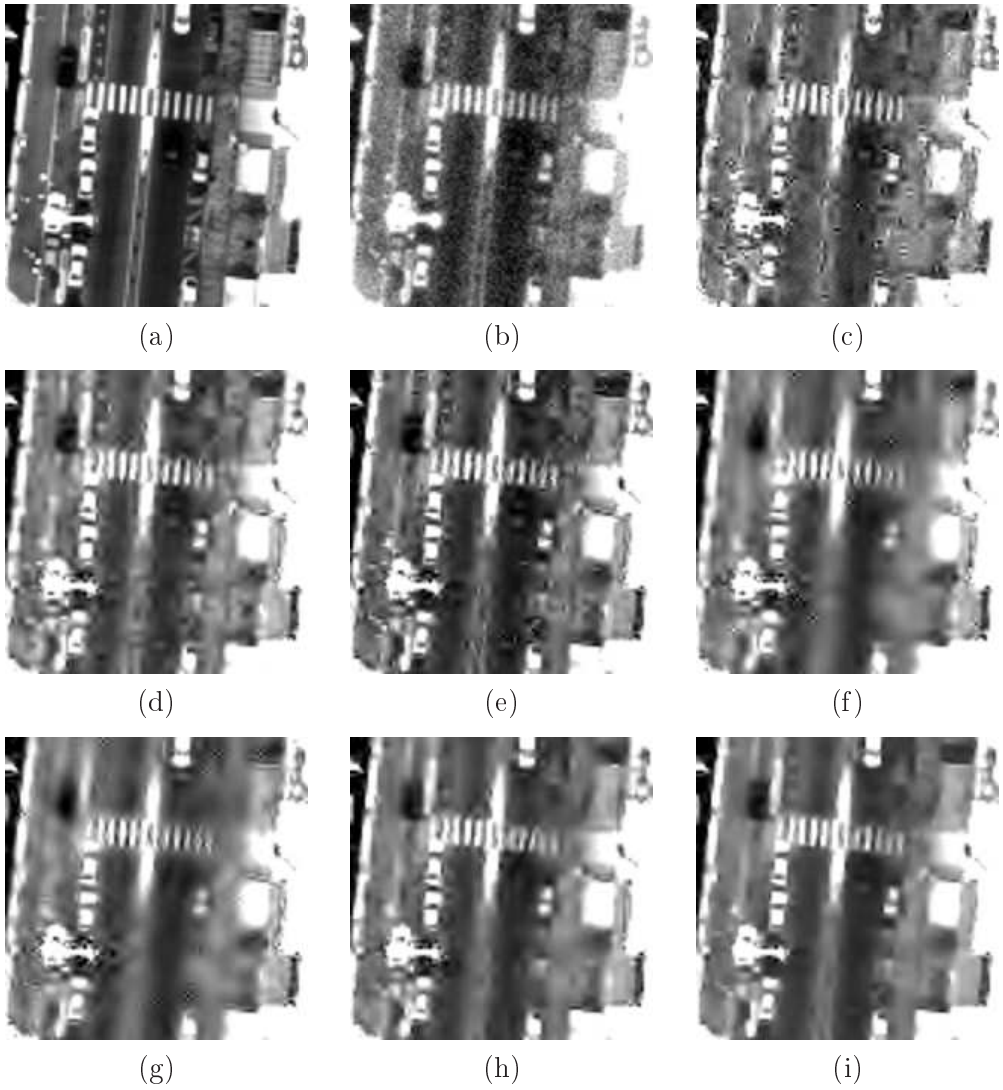


Figure 5.10: Visual comparison of on-board and on-ground chains. Displayed images have a size of 200×200 pixels. (a) is the reference image, (b) is the instrumental image (output of the acquisition, $PSNR = 32.69\ dB$), (c) is the reconstructed image provided by the CNES ($PSNR = 45.93\ dB$), (d) and (e) are the reconstructed images respectively from the Shearlets based on-board ($PSNR = 46.80\ dB$) and on-ground ($PSNR = 46.69\ dB$) chains, (f) and (g) are the reconstructed images respectively from the block thresholding based on-board ($PSNR = 46.46\ dB$) and on-ground ($PSNR = 46.24\ dB$) chains, (h) and (i) are the reconstructed images respectively from the ForWarRD based on-board ($PSNR = 47.11\ dB$) and on-ground ($PSNR = 47.05\ dB$) chains. The target rate is 2.5 bits/pixel and the simulated SNR is 30-100. The image range has been extended to point up the image reconstruction artifacts.

as taking the original wavelet coefficients summed with negative impulses (where the magnitude of the impulses is equal to the value of the coefficients prior to quantizing). The inverse transform, performed after the transmission, displays then the wavelet responses.

These artifacts visually look like checkerboard (see Fig. 5.11) and are thus sometimes referred that way in the literature [Selesnick 2003]. Clearly, these structures are not appreciated in the final image as they can not be related to some natural image features. The denoising of such coding noise is then important for the quality of the final image and is the focus of this part.

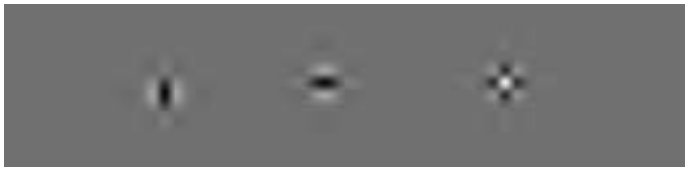


Figure 5.11: Wavelet responses for the first level of a 3-levels CDF 9/7 decomposition. The first two wavelets are oriented in the vertical and horizontal directions. The third wavelet is a mix of two diagonal orientations and gives the "checkerboard" artifact.

We start by giving in this section a brief review of the state-of-the-art of quantization noise removal methods. We will then discuss in Section 5.3 the integration of these techniques in the satellite imaging chain.

5.2.1 Variational methods for denoising quantization noise

Several methods have been recently proposed in [Durand 2003, Weiss 2008, Tramini 1998] to tackle the problem of quantization noise removal for wavelet-based coder. They proposed to solve the problem of retrieving an image x_0 from its coded version $\tilde{x}$. The observed coded image $\tilde{x}$ can be modeled as

$$\tilde{x} = \tilde{W}\left(Q\left(Wx_0\right)\right), \tag{5.5}$$

where W stands for a wavelet transform (its inverse is denoted $\tilde{W}$) and Q is a quantizing process. Techniques [Weiss 2008] and [Tramini 1998] are actually very similar and, consequently, we only present the methods proposed in [Durand 2003] and [Weiss 2008]. These methods are both based on a variational framework and both rely on the minimization of the total variation (TV) prior [Rudin 1992].

The TV prior assumes that an image can be modeled as a smooth function with discontinuities across curves. The oscillations created by the coding artifacts cannot therefore be considered to be natural and do not belong to an image. The particularity of these artifacts is that they exhibit important variations of intensity which tend to increase the magnitude of the gradient of the image, assumed to be low by the smoothness hypothesis. Minimizing the l^1-norm of the gradient of the image, namely the TV, will then replace these oscillations by smooth homogeneous

regions. Both methods [Durand 2003] and [Weiss 2008] could globally be formalized as the following minimization problem

$$\hat{x} = \begin{array}{ll} \arg\min & \|\nabla x\|_1 \\ \text{subject to} & x \in K \end{array}, \tag{5.6}$$

where $\hat{x}$ is the denoised image and K is a set that constrains the reconstructed image. Two different approaches have been proposed in [Durand 2003] and [Weiss 2008] to formulate this set. The authors of [Weiss 2008] proposed to define the set K such that it constrains the error between the observed and the reconstructed wavelet coefficients. In detail, let $\mathcal{Q}$ be the set of all possible output quantized values $\mathcal{Q} = \{q_k; k \in \mathbb{Z}, q_0 = 0\}$ and b_k, b_{k+1} $(b_{k+1} > b_k)$ be the boundaries of each quantization interval such that

$$(W\tilde{x})_i = q_k, \quad \text{if } b_k \leq (Wx_0)_i < b_{k+1}, \quad \forall i \in \{0, \ldots, N-1\}. \tag{5.7}$$

From equation (5.7), we have

$$b_k - q_k \leq (Wx_0)_i - (W\tilde{x})_i < b_{k+1} - q_k, \quad \forall i \in \{0, \ldots, N-1\}. \tag{5.8}$$

For each pixel i, we set the bounds $\alpha_i = b_k - q_k$ and $\beta_i = b_{k+1} - q_k$, where k verifies (5.8) given i. Note that the bounds α_i and β_i can be estimated from the wavelet coefficients of the decoded image and the knowledge of the quantizing model. The authors of [Weiss 2008] proposed to define K as the following hypercube

$$K = \left\{x \in \mathbb{R}^N, \alpha_i \leq (Wx)_i - (W\tilde{x})_i < \beta_i, \forall i \in \{0, \ldots, N-1\}\right\}, \tag{5.9}$$

such that problem (5.6) consists in minimizing the TV of the image under the constraint that the error between the wavelet coefficients of the reconstructed image and the wavelet coefficients of the decoded image belongs to the invervals defined by the boundaries (5.8).

The method proposed in [Durand 2003] is slightly different and constrains the wavelet coefficients without any reference to the original image x_0. They define the set K as

$$K = \left\{x \in \mathbb{R}^N, (Wx)_i = (W\tilde{x})_i, \forall i \in M\right\}, \tag{5.10}$$

where M is the set of coefficients coordinates that have not been set to zero by the quantizing

$$M = \left\{i \in \{0, \ldots, N-1\}, |(W\tilde{x})_i| > 0\right\}. \tag{5.11}$$

The idea of the method proposed in [Durand 2003] is to reconstruct the small coefficients that have been set to zero by the quantizing. The method relies on the fact that the minimization of the TV creates flat regions which are represented by small wavelet coefficients. The presence of the constraint (5.10) is to ensure that only these small coefficients are updated and that the large quantized coefficients, which are likely to be close to the original ones, remain unchanged.

A comparison of the two presented methods is given at the end of this part. We will see however that the flat homogeneous regions created by the minimization of the TV are not natural in the sense that they cannot be interpreted as some physical features of an image. The problem of quantization noise removal is actually very difficult to address. The main difficulty lies in the fact that the quantization noise is highly correlated to the signal source and cannot be modeled using classical probability distributions (except under high coding rate assumption). We present in the next part methods to improve the statistical properties of the quantization noise.

5.2.2 Dithering methods for removing quantization artifacts

We present in this part dithering techniques to reduce the quantization artifacts. These techniques have been originally introduced in the speech [Jayant 1972] and video [Roberts 1962] processing communities to reduce the perceptual distortion due to compression. The particularity of these techniques is that they consist in inserting a noise prior to quantizing to improve the statistics of the quantization error. A review of the theory of dithering techniques is given in Appendix B.

For the application of quantization artifacts, we will focuss here on the subtractive dithering system proposed in [Schuchman 1964] whose particularity is to subtract the added noise after quantizing. From Appendix B, we see that the non-subtractive dithering technique only allows the moments of the global error ε to be decorrelated to the source w. An independence of the moments is however rarely exploited by restoration algorithms, which require the true signal independence, only provided by the subtractive variant. Let w be an original (i.e. prior to quantizing) wavelet subband and $\tilde{w}$ be the output corresponding subband which, for a subtractive dithering system, writes

$$\tilde{w} = Q(w + v) - v, \tag{5.12}$$

where Q is the quantizing operator and v is the dithering noise. The global error ε of this sytem is defined as

$$\varepsilon = \tilde{w} - w. \tag{5.13}$$

As mentioned by [Lipshitz 1992], a subtractive dithering system produces an independent and uniformly distributed global error if the dithering noise v can be expressed as the summation of rectangular probability density functions. This is an encouraging result as it implies that an on-board restoration coupled with a subtractive dithering scheme will result in a restored image with a residual noise which is independent of the original image. Since this residual noise is not structured, it can be interpreted physically (as the intrumental noise of the sensor for example) better than the residual noise obtained with the current imaging chain system. This aspect of residual noise is very important as it is one of the features seeked by the CNES for the design of restoration methods [Dherete 2003]. We will discuss this aspect later as this is the basis of the proposed imaging chain described in Section 5.3.

We would like also to mention the dithering technique proposed in [Stamm 2011]. This method is slightly different from the dithering techniques presented in Appendix B as it is more focused on the reconstruction of the original wavelet subbands rather than improving the statistics of the quantization noise. More precisely, the main result of [Stamm 2011] states that the probability density function of a wavelet subband can be recovered exactly (assuming we know the parameters of its model) from its quantized version by adding a dithering noise v to the quantized coefficients.

We assume that the quantizing model is the same than the one presented in Section 5.2.1. The authors of [Stamm 2011] proposed to model a wavelet subband w (each subband can be treated separately) by a Laplace distribution [Li 1998]

$$p_w(w) = \frac{\lambda}{2} e^{-\lambda|w|}, \tag{5.14}$$

where λ is the scale parameter that can be estimated using classical estimation techniques such as least-squares minimization methods or maximum-likelihood estimations. Similarly to (5.7), the quantized wavelet subband $\tilde{w}$ writes

$$\tilde{w} = q_k, \quad \text{if } b_k \leq w < b_{k+1}. \tag{5.15}$$

Using the wavelet subband model (5.14), we can express the probability density function $p_{\tilde{w}}$ of a quantized wavelet subband

$$p_{\tilde{w}}(\tilde{w} = q_k) = \begin{cases} \frac{1}{2}\left(e^{-\lambda b_k} - e^{-\lambda b_{k+1}}\right), & \text{if } k \geq 1 \\ 1 - \frac{1}{2}\left(e^{-\lambda b_0} - e^{-\lambda b_1}\right), & \text{if } k = 0 \\ \frac{1}{2}\left(e^{\lambda b_{k+1}} - e^{\lambda b_k}\right), & \text{if } k \leq -1. \end{cases} \tag{5.16}$$

As said previously, the method proposed in [Stamm 2011] consists in adding a dithering noise after quantizing. The final wavelet subband z is then given by

$$z = \tilde{w} + v, \tag{5.17}$$

where $\tilde{w}$ is the quantized wavelet subband and v the dithering noise. The wavelet subband probability density function p_z can be expressed using the law of total probability [Stamm 2011]

$$p_z(z) = \sum_{k=-\infty}^{+\infty} p_{z|\tilde{w}}(z|\tilde{w} = q_k) p_{\tilde{w}}(\tilde{w} = q_k), \tag{5.18}$$

where

$$p_{z|\tilde{w}}(z|\tilde{w} = q_k) = p_{v|\tilde{w}}(v = z - q_k|\tilde{w} = q_k), \tag{5.19}$$

is the probability density function of the dither noise v knowing the quantized values $\tilde{w}$. The authors of [Stamm 2011] showed that the choices

$$p_{v|\tilde{w}}(v|\tilde{w}=q_k, k\neq 0) = \begin{cases} \frac{1}{\alpha_k} e^{-\operatorname{sign}(q_k)\hat{\lambda}v}, & \text{if } (b_k - q_k) \leq v < (b_{k+1} - q_k) \\ 0, & \text{otherwise} \end{cases} \tag{5.20}$$

$$p_{v|\tilde{w}}(v|\tilde{w}=0) = \begin{cases} \frac{1}{\alpha_0} e^{-\hat{\lambda}|v|}, & \text{if } b_0 > v > b_1 \\ 0, & \text{otherwise} \end{cases}, \tag{5.21}$$

with a_k being some normalization constants and $\hat{\lambda}$ an estimated value of the scale parameter λ, lead to the original wavelet subband probability density function p_w, under the condition that the scale parameter has been estimated exactly, i.e. $\hat{\lambda} = \lambda$ [Stamm 2011]

$$\begin{aligned} p_z(z) &= \sum_{k=-\infty}^{+\infty} p_{z|\tilde{w}}(z|\tilde{w}=q_k) p_y(\tilde{w}=q_k) \\ &= \sum_{k=-\infty}^{-1} \frac{1}{\alpha_k} e^{\hat{\lambda}(z-q_k)} \frac{1}{2} \left(e^{\hat{\lambda} b_{k+1}} - e^{\hat{\lambda} b_k} \right) \mathbb{1}(b_k \leq z < q_{k+1}) \\ &+ \frac{1}{\alpha_0} e^{-\hat{\lambda}|z|} \left(1 - \frac{1}{2} \left(e^{\hat{\lambda} b_0} - e^{-\hat{\lambda} b_1} \right) \right) \mathbb{1}(b_0 \leq z < b_1) \\ &+ \sum_{k=1}^{+\infty} \frac{1}{\alpha_k} e^{-\hat{\lambda}(z-q_k)} \frac{1}{2} \left(e^{-\hat{\lambda} b_k} - e^{-\hat{\lambda} b_{k+1}} \right) \mathbb{1}(b_k \leq z < q_{k+1}) \\ &= \frac{\lambda}{2} e^{-\hat{\lambda}|z|} = p_w(z), \end{aligned} \tag{5.22}$$

where

$$\mathbb{1}(a \leq z < b) = \begin{cases} 1, & \text{if } a \leq z < b \\ 0, & \text{otherwise} \end{cases}. \tag{5.23}$$

Even if the reconstructed and original subbands will numerically differ, this technique will remove the undesirable observed artifacts, due to the quantization, by filling in the blanks. The fact that we also add dither noise on the null coefficients may also provide the residual noise appreciated by image analysis experts.

5.2.3 Comparison of removal methods for quantization artifacts

We simulate the behavior of the presented quantization removal methods directly on a coded version of the reference (i.e. without any blur or instrumental noise) satellite image shown Fig. 5.3. The simulation of the complete imaging chain including these techniques is done in the next part. To perform a fair comparison, the image will be coded using the biorthogonal 9/7 wavelet transform [Cohen 1992] followed by the quantizer described in [Lipshitz 1992]. As a consequence, the method [Stamm 2011] has been adapted to this choice. For the subtractive dithering method

[Lipshitz 1992], we simulated a uniform dithering noise to limit the power of the residual noise. This dithering noise will be applied to the wavelet subbands of the image prior to quantizing. Therefore, after the inverse transform the residual noise (i.e. the error between the reference image and the output of the dithering system) is not uniformly distributed anymore but we found out experimentaly that this noise appears, suprisingly, to be still independent and identically distributed following a centered Gaussian law.

We only provide visual results as common criteria such as PSNR do not take into account the appreciated physical perception of residual noise.

The results are given Fig. 5.12 to 5.15. Visually, we immediately see that the techniques based on the minimization of the TV create large smooth homogenous regions and remove the small details of the image. This effect is known as the *cartoon* effect. These flat regions are not considered to be natural for a satellite image and are really not appreciated by image analysis experts who clearly prefer a deterioration that can be interpreted physically. As explained previously, this is for example the case of an unstructured residual noise. The subtractive dithering technique and the method proposed in [Stamm 2011] give good visual results in this sense. Both images are well reconstructed and do not present common artifacts such as ringing or blurry edges. The quality of the image reconstructed with the subtractive dithering technique actually seems slightly better, particularly on the small details of the image (cars and zebras). As expected, these methods leave a residual noise on the reconstructed image which can be interpretated as the instrumental noise of the sensor.

5.3 Proposed imaging chain

In the previous section, we showed that the dithering techniques may be very interesting to remove the structured artifacts of the coding step. As we have also mentioned in Section 5.2.3, these techniques leave a uniform residual noise which is highly appreciated from the image analysis experts as it can be interpretated physically. More precisely, an ideal restored image (as defined by image analysis experts) should owns a residual blur characterized by a target PSF [Lambert-Nebout 2000] along with a uniform residual noise with a fixed standard deviation [Dherete 2003].

We also presented in Section 5.1.1 an on-board restoration technique which gives an image with a residual noise (whose power is very small in comparison to the power of the residual noise obtained from the dithering techniques) and a residual blur fully characterized by the target PSF. If we combine these two techniques, i.e. if we use an on-board restoration coupled with a subtractive dithering technique, the image obtained at the output of the chain will then present an unstructured residual noise (coming from the dithering technique) with the blur of the target PSF (coming from the on-board restoration). And as mentioned previously, a final image with such characteristics is the objective of image analysis experts as it can be interpreted as the direct output of an ideal instrument.

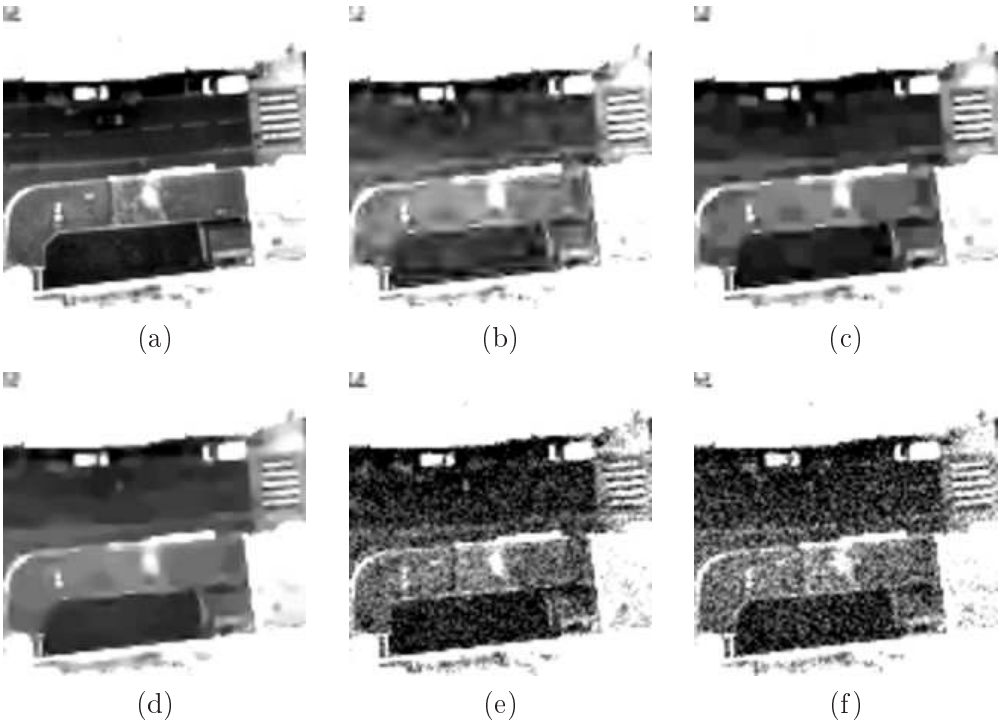


Figure 5.12: Visual comparison of quantizing removal techniques. Displayed images have a size of 200×200 pixels. (a) is the reference image, (b) is the decompressed image, (c) is the image obtained using the post-processing technique proposed in [Durand 2003], and (d) is the image obtained using the post-processing technique proposed in [Weiss 2008], (e) is the image reconstructed using the post-processing dithering technique proposed in [Stamm 2011], (f) is the image reconstructed using the subtractive dithering technique [Lipshitz 1992] with an uniform dithering noise. The target rate is 2.5 bits/pixel. The image range has been extended to point up the image reconstruction artifacts.

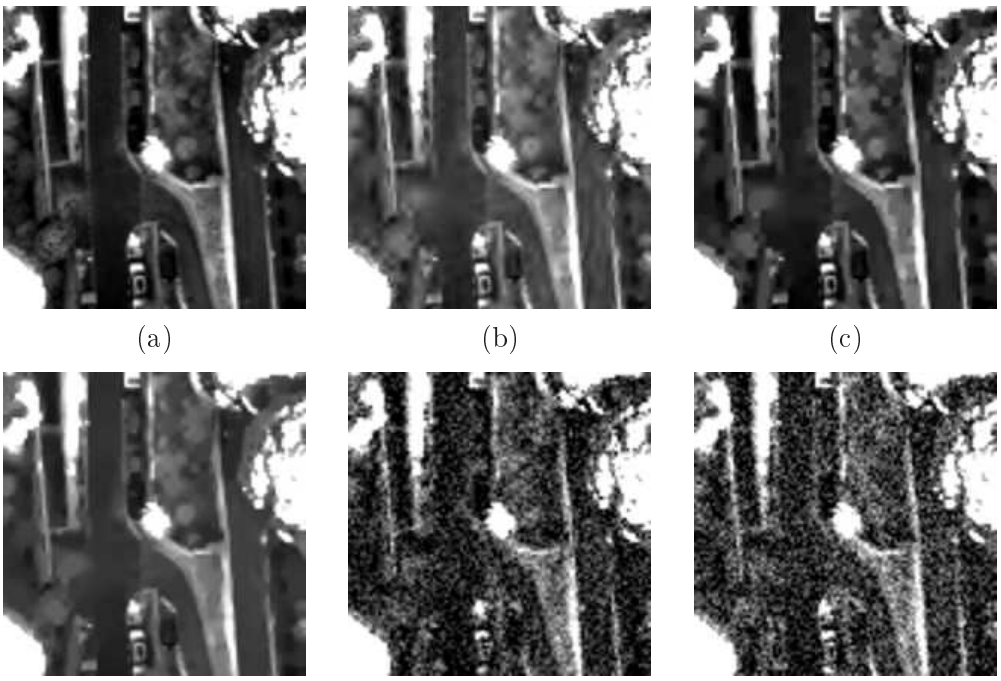


Figure 5.13: Visual comparison of quantizing removal techniques. Displayed images have a size of 200 × 200 pixels. (a) is the reference image, (b) is the decompressed image, (c) is the image obtained using the post-processing technique proposed in [Durand 2003], and (d) is the image obtained using the post-processing technique proposed in [Weiss 2008], (e) is the image reconstructed using the post-processing dithering technique proposed in [Stamm 2011], (f) is the image reconstructed using the subtractive dithering technique [Lipshitz 1992] with an uniform dithering noise. The target rate is 2.5 bits/pixel. The image range has been extended to point up the image reconstruction artifacts.

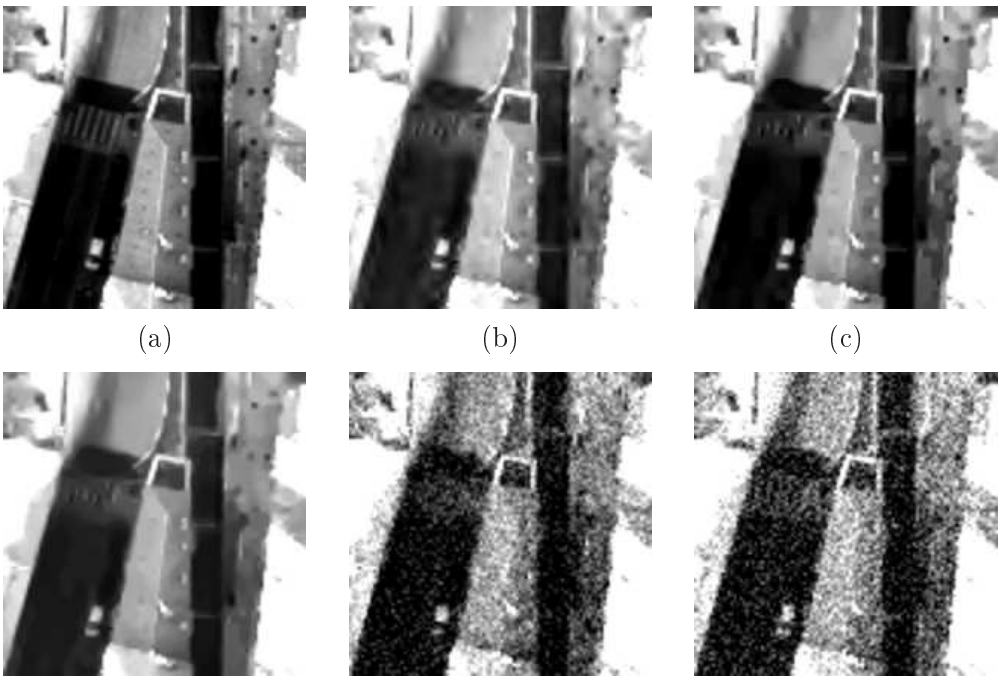


Figure 5.14: Visual comparison of quantizing removal techniques. Displayed images have a size of 200×200 pixels. (a) is the reference image, (b) is the decompressed image, (c) is the image obtained using the post-processing technique proposed in [Durand 2003], and (d) is the image obtained using the post-processing technique proposed in [Weiss 2008], (e) is the image reconstructed using the post-processing dithering technique proposed in [Stamm 2011], (f) is the image reconstructed using the subtractive dithering technique [Lipshitz 1992] with an uniform dithering noise. The target rate is 2.5 bits/pixel. The image range has been extended to point up the image reconstruction artifacts.

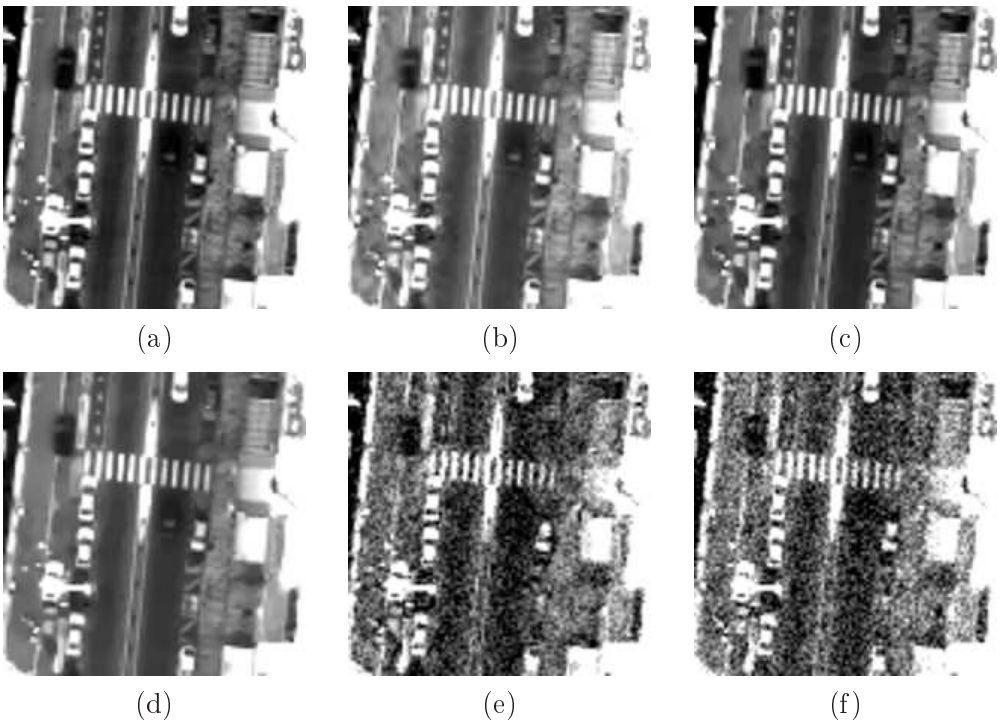


Figure 5.15: Visual comparison of quantizing removal techniques. Displayed images have a size of 200 × 200 pixels. (a) is the reference image, (b) is the decompressed image, (c) is the image obtained using the post-processing technique proposed in [Durand 2003], and (d) is the image obtained using the post-processing technique proposed in [Weiss 2008], (e) is the image reconstructed using the post-processing dithering technique proposed in [Stamm 2011], (f) is the image reconstructed using the subtractive dithering technique [Lipshitz 1992] with an uniform dithering noise. The target rate is 2.5 bits/pixel. The image range has been extended to point up the image reconstruction artifacts.

From this remark, we propose the imaging chain shown Fig. 5.16.

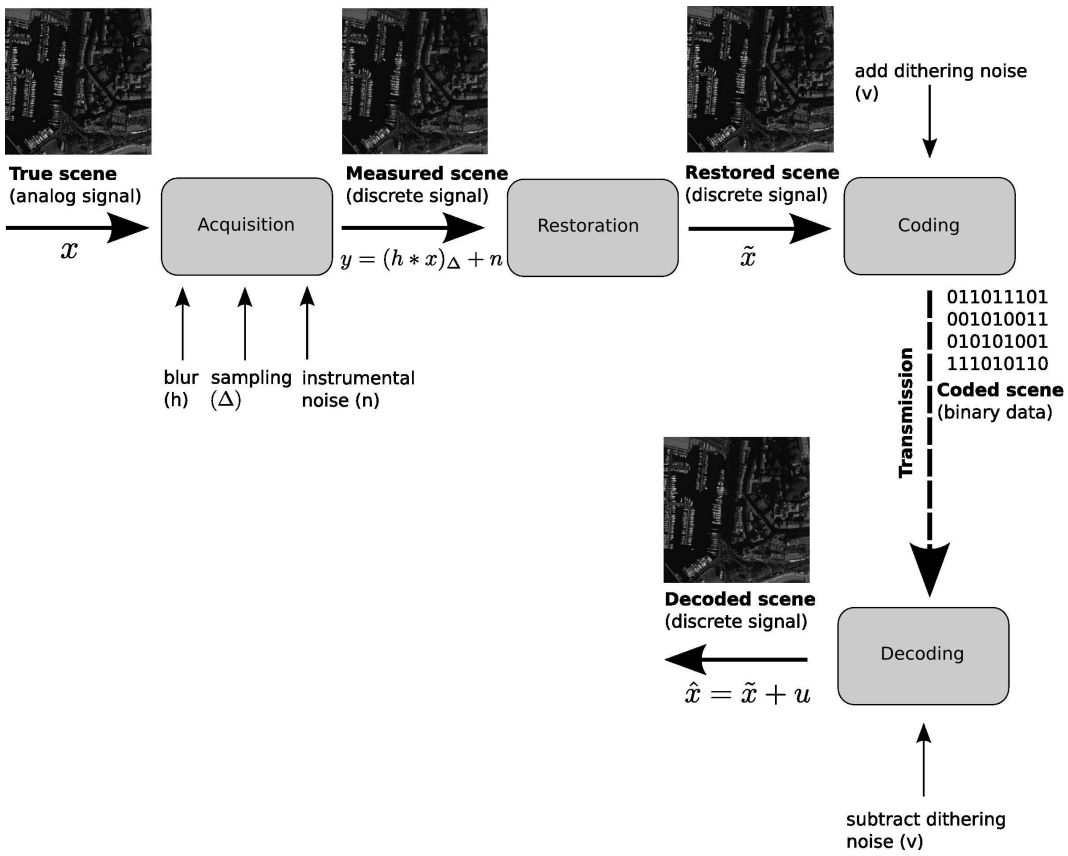
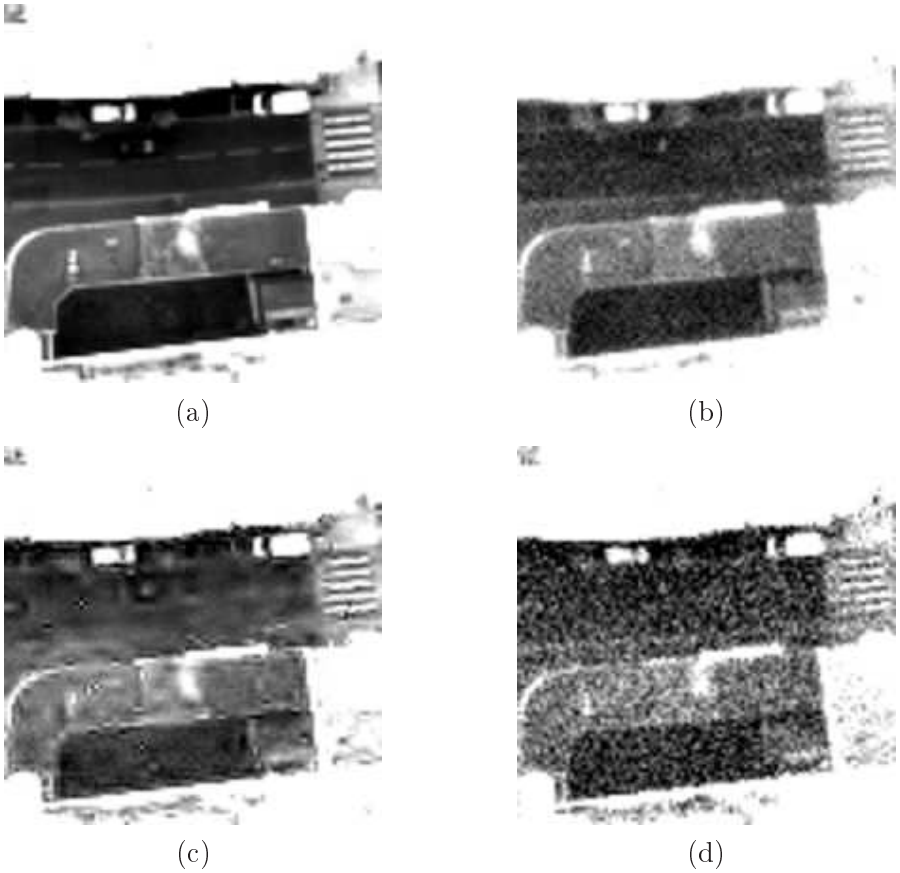


Figure 5.16: Proposed satellite imaging chain

This chain includes the on-board restoration based on the Shearlets transform [Patel 2009] and the subtractive dithering technique [Lipshitz 1992] to decorrelate the quantizing noise. Note that, in this chain, the quantizer follows the model described in [Lipshitz 1992] to respect the subtractive dithering scheme hypothesis. The coding step is then decomposed in a 3-levels CDF 9/7 wavelet transform followed by an explicit quantization of the wavelet coefficients and an entropy encoding of the quantized coefficients. The results of the proposed imaging chain are given Fig. 5.17 to 5.20.

We immediately see that the reconstructed images with the proposed chain do not present any common wavelet compression artifacts (see figures 5.17 and 5.18), that we observed on the reconstructed image provided by the CNES. They exhibit instead an unstructured residual noise which is visually similar to the noise obtained on the instrumental image at the ouput of the acquisition chain. This is particularly visible on the dark zones of the reconstructed image, see figures 5.18 and 5.19.

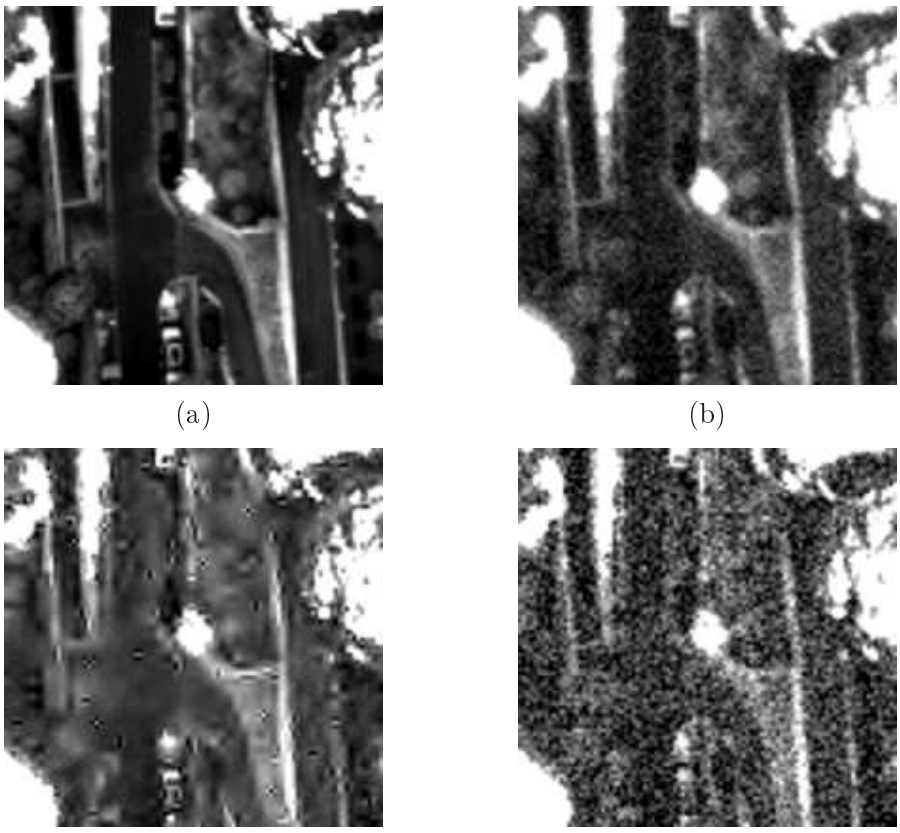
It is clear that the proprosed chain tends to replace one type of residual noise (wavelet compression artifacts) by another one. The obtained residual noise is how-



(a) (b)

(c) (d)

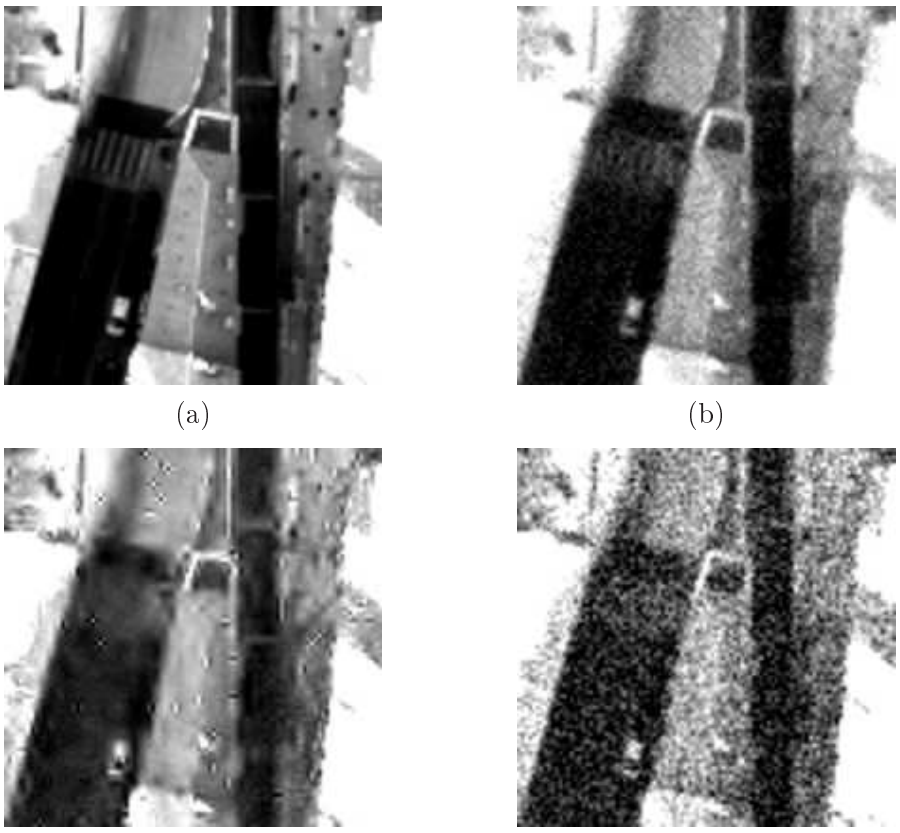
Figure 5.17: Visual comparison of the proposed and the current imaging chains. Displayed images have a size of 200×200 pixels. (a) is the reference image, (b) is the instrumental image, (c) is the decompressed and restored image provided by the CNES, (d) is the reconstructed image from the Shearlets based on-board chain followed by a subtractive dithering scheme. The target rate is 2.5 bits/pixel and the simulated SNR is 30-100. The image range has been extended to point up the image reconstruction artifacts.



(a) (b)

(c) (d)

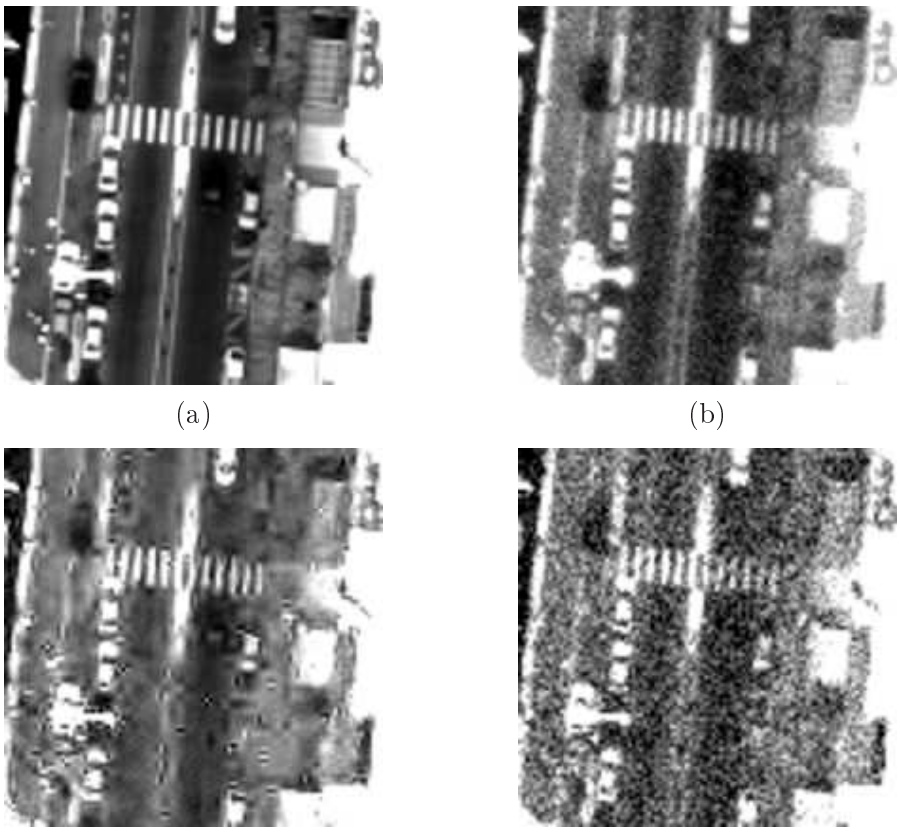
Figure 5.18: Visual comparison of the proposed and the current imaging chains. Displayed images have a size of 200×200 pixels. (a) is the reference image, (b) is the instrumental image, (c) is the decompressed and restored image provided by the CNES, (d) is the reconstructed image from the Shearlets based on-board chain followed by a subtractive dithering scheme. The target rate is 2.5 bits/pixel and the simulated SNR is 30-100. The image range has been extended to point up the image reconstruction artifacts.



(a) (b)

(c) (d)

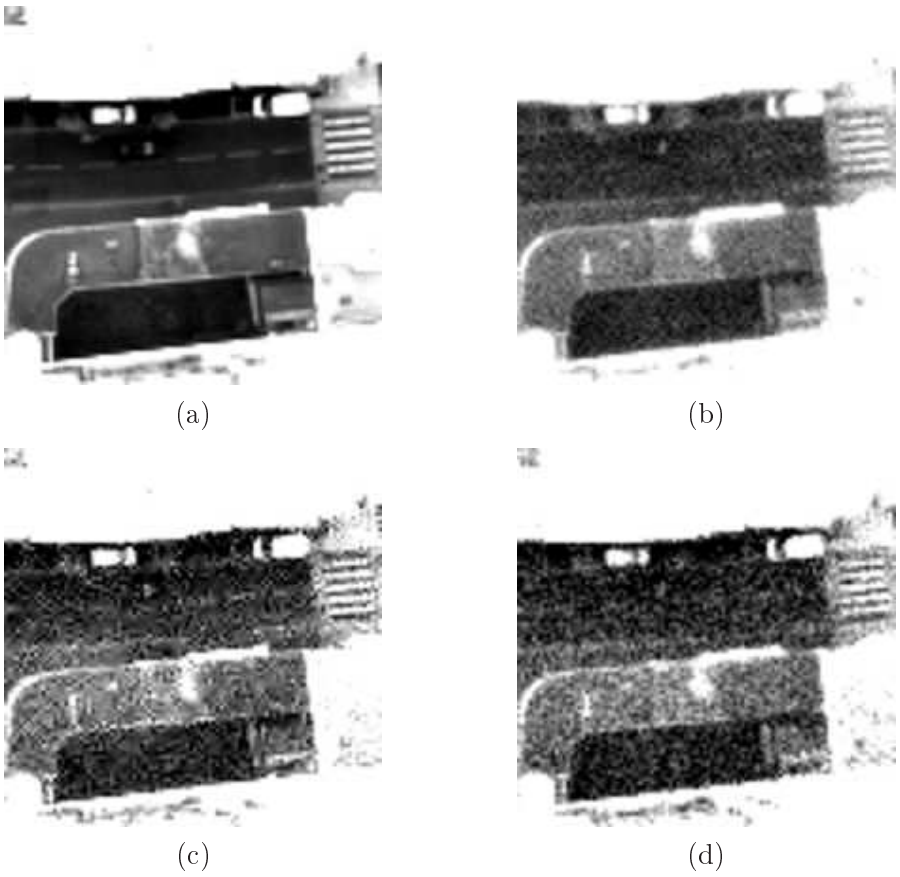
Figure 5.19: Visual comparison of the proposed and the current imaging chains. Displayed images have a size of 200×200 pixels. (a) is the reference image, (b) is the instrumental image, (c) is the decompressed and restored image provided by the CNES, (d) is the reconstructed image from the Shearlets based on-board chain followed by a subtractive dithering scheme. The target rate is 2.5 bits/pixel and the simulated SNR is 30-100. The image range has been extended to point up the image reconstruction artifacts.



(a) (b)

(c) (d)

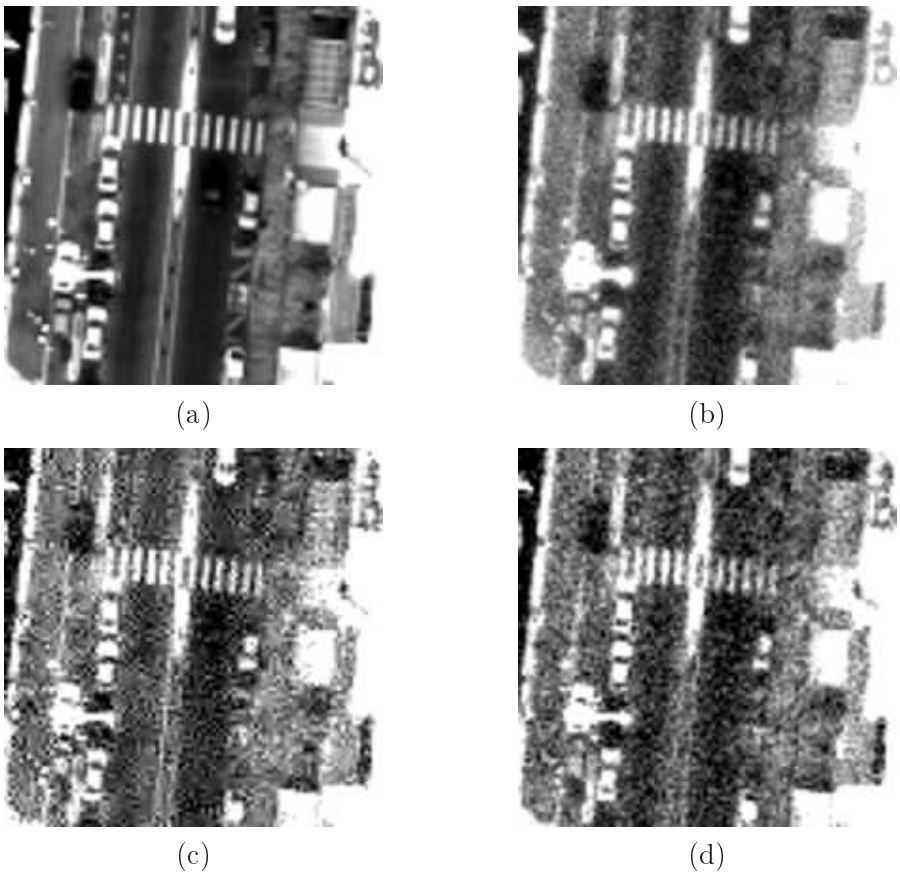
Figure 5.20: Visual comparison of the proposed and the current imaging chains. Displayed images have a size of 200×200 pixels. (a) is the reference image, (b) is the instrumental image, (c) is the decompressed and restored image provided by the CNES, (d) is the reconstructed image from the Shearlets based on-board chain followed by a subtractive dithering scheme. The target rate is 2.5 bits/pixel and the simulated SNR is 30-100. The image range has been extended to point up the image reconstruction artifacts.



(a) (b)

(c) (d)

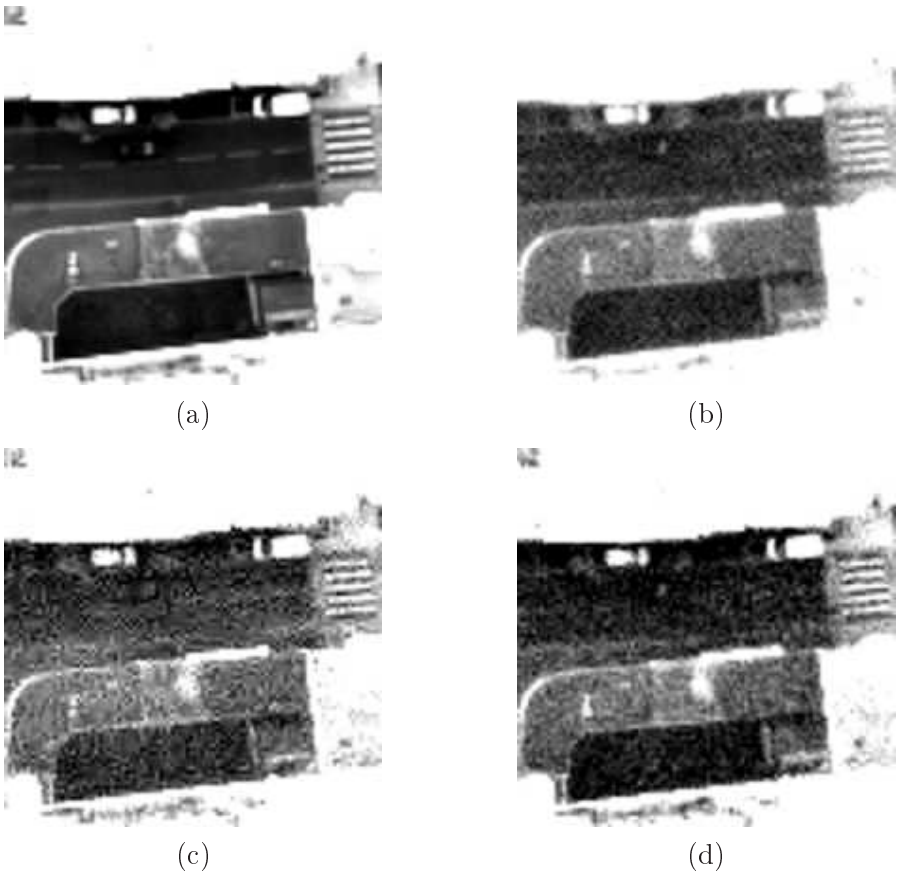
Figure 5.21: Visual comparison of the proposed and the current imaging chains. Displayed images have a size of 200×200 pixels. (a) is the reference image, (b) is the instrumental image, (c) is the decompressed and restored image provided by the CNES, (d) is the reconstructed image from the Shearlets based on-board chain followed by a subtractive dithering scheme. The target rate is 3.0 bits/pixel and the simulated SNR is 30-100. The image range has been extended to point up the image reconstruction artifacts.



(a) (b)

(c) (d)

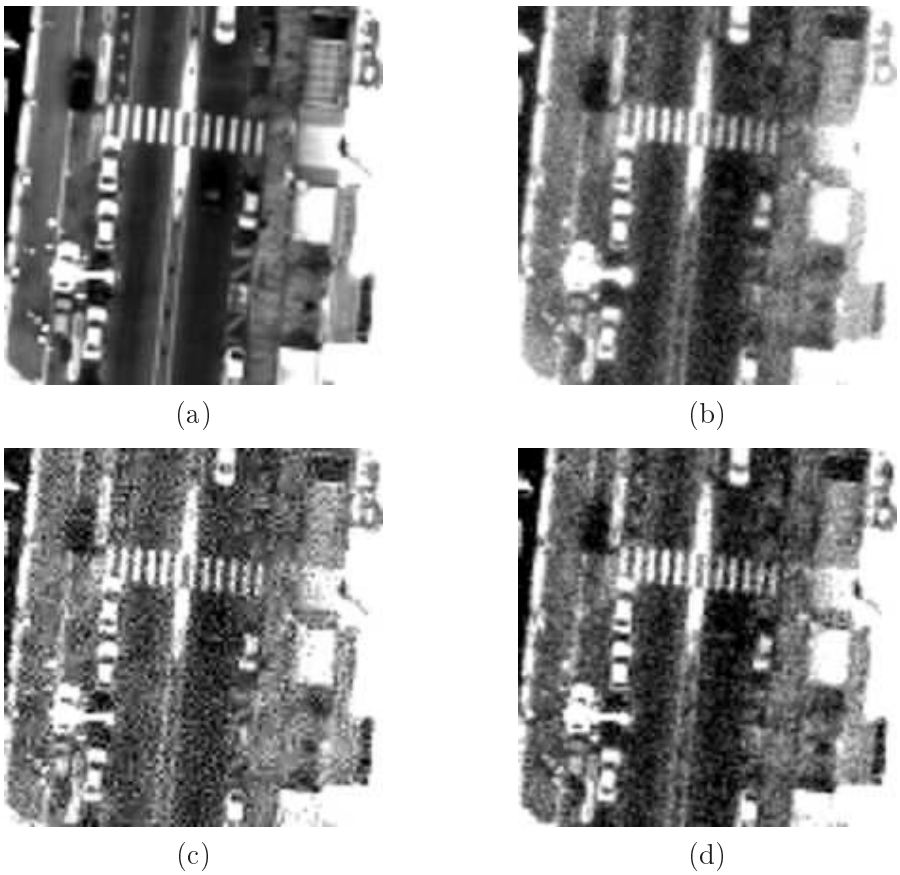
Figure 5.22: Visual comparison of the proposed and the current imaging chains. Displayed images have a size of 200×200 pixels. (a) is the reference image, (b) is the instrumental image, (c) is the decompressed and restored image provided by the CNES, (d) is the reconstructed image from the Shearlets based on-board chain followed by a subtractive dithering scheme. The target rate is 3.0 bits/pixel and the simulated SNR is 30-100. The image range has been extended to point up the image reconstruction artifacts.



(a) (b)

(c) (d)

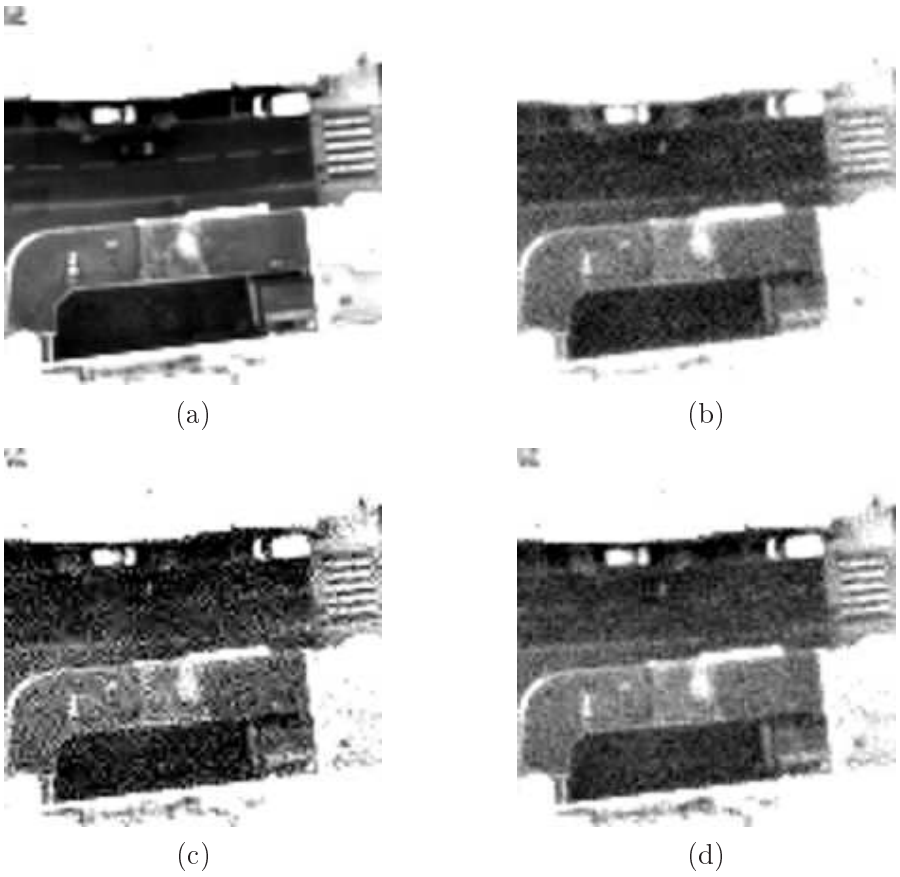
Figure 5.23: Visual comparison of the proposed and the current imaging chains. Displayed images have a size of 200×200 pixels. (a) is the reference image, (b) is the instrumental image, (c) is the decompressed and restored image provided by the CNES, (d) is the reconstructed image from the Shearlets based on-board chain followed by a subtractive dithering scheme. The target rate is 3.5 bits/pixel and the simulated SNR is 30-100. The image range has been extended to point up the image reconstruction artifacts.



(a) (b)

(c) (d)

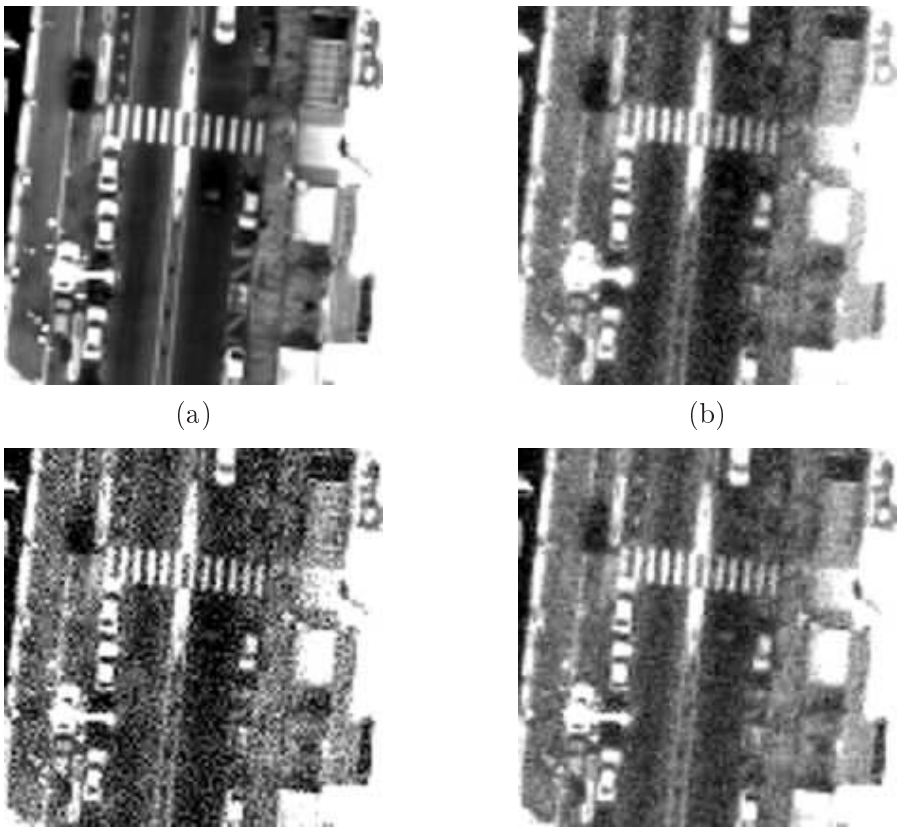
Figure 5.24: Visual comparison of the proposed and the current imaging chains. Displayed images have a size of 200×200 pixels. (a) is the reference image, (b) is the instrumental image, (c) is the decompressed and restored image provided by the CNES, (d) is the reconstructed image from the Shearlets based on-board chain followed by a subtractive dithering scheme. The target rate is 3.5 bits/pixel and the simulated SNR is 30-100. The image range has been extended to point up the image reconstruction artifacts.



(a) (b)

(c) (d)

Figure 5.25: Visual comparison of the proposed and the current imaging chains. Displayed images have a size of 200×200 pixels. (a) is the reference image, (b) is the instrumental image, (c) is the decompressed and restored image provided by the CNES, (d) is the reconstructed image from the Shearlets based on-board chain followed by a subtractive dithering scheme. The target rate is 4.0 bits/pixel and the simulated SNR is 30-100. The image range has been extended to point up the image reconstruction artifacts.



(a) (b)

(c) (d)

Figure 5.26: Visual comparison of the proposed and the current imaging chains. Displayed images have a size of 200×200 pixels. (a) is the reference image, (b) is the instrumental image, (c) is the decompressed and restored image provided by the CNES, (d) is the reconstructed image from the Shearlets based on-board chain followed by a subtractive dithering scheme. The target rate is 4.0 bits/pixel and the simulated SNR is 30-100. The image range has been extended to point up the image reconstruction artifacts.

ever better appreciated by image analysis experts as it can be interpreted physically. More precisely, the proprosed imaging chain produces a reconstructed image which owns the two characteristics of an ideal image: Blur with the target PSF (obtained by the on-board restoration) and a residual unstructured noise [Dherete 2003]. The drawback of the proposed method is that the standard deviation of the residual noise is function of the quantizing step (see Theorem 11) while it should be constant for all coding rates. Consequently, for a low coding rate, the proposed chain gives an image which is more noisy than the instrumental one. It gives however very interesting results for high coding rates as shown by Fig. 5.21 to 5.26. Futher works need thus to be done on this aspect.

5.4 Conclusions and perspectives

In this chapter, we presented a numerical study on the satellite imaging chain optimization problem. We presented several results which showed that the quality of the reconstructed image can be improved if one concedes several changes on the usual design of imaging chains.

The first one would be to move the restoration step on-board of satellite, prior to coding. The results we obtained here showed that an on-board restoration allows to reconstruct an image with less reconstruction artifacts, specially on shadows zones. On a more theoretical point of view, moving the restoration on-board seems to be a reliable method to approximately optimize the global imaging chain since it does not require to express the global distortion as a function of the parameters of the chain which, as discussed in Chapter 4.4, is difficult for true satellite imaging systems.

The second point discussed in this chapter deals with the problem of coding noise removal. From the results we presented, we concluded that the current state-of-the-art coding noise denoising algorithms do not give competitve results and that the best option may be to used dithering techniques to transform the stuctured coding noise in an unstructured residual noise. This property of residual noise is highly appreciated from photo interpreters since it simulates the noise obtained directly at the output of the instrument. From these conclusions, we proposed a new imaging chain based on an on-board restoration coupled with a subtractive dithering technique. We showed results on a real satellite data and we compared the results of the proposed chain with the ones obtained with the current satellite imaging chain used by the CNES. We showed that the proposed chain gives interesting results and may be particularly efficient at medium and high coding rates (around 3.0 bits/pixel and more). The particularity of the proposed imaging chain is that the final image is fully characterized by the target blur (specified by the CNES) and a residual unstructured noise. Such feature is interesting for images analysis experts since classical defects of the compression and restoration steps do not appear in the final image, such that these two steps appear then almost transparent in the chain.

A drawback of the proposed method is that the power of this residual noise depends on the target coding rate. At low coding rate (like 2.5 bits/pixel), the

final image appears to be more noisy than the instrumental image and is therefore difficult to exploit. It would be thus interesting to investigate how to limit the intensity of this residual noise such that competitive results can also be obtained at low coding rates.

Part III

A compressed sensing based satellite imaging chain

Chapter 6

Compressed Sensing for satellite imaging

The last part of the thesis is dedicated to the study of Compressed Sensing (CS) for satellite imaging. This study on the Compressed Sensing slightly differs from the global optimization techniques that we presented in part II and the purpose of this study is mainly to evaluate the capability of CS applied to high resolution satellite imaging. The CS technique is interesting for satellite imaging as it simplifies the resources required for the acquisition of the image, which are, in our case performed on-board of the satellite. All the processings on the image are then performed on-ground by a specific decoder and the quality of the final image entirely depends on the reliability of this decoder. Due to the limited capacity of embedded resources, this technique clearly appears to be adapted to our context.

We first present in Section 6.1 a brief introduction of the CS framework. We then detail, in Section 6.2, how to apply this technique to satellite imging. We present reconstruction results of the proposed method in comparison to the results obtained with the current imaging chain and we conclude this part.

6.1 A short introduction to Compressed Sensing

6.1.1 Motivations

In a classical imaging system, the acquired image is sampled at the Nyquist frequency to give N pixels. Any digital camera produces nowadays an image with dozen millions of pixels. By assuming that each pixel is represented on 24 bits (8 bits per color channel), each image requires then almost 100 Mb of storage capacity. Some compression algorithms, like the JPEG [Wallace 1992] and JPEG2000 standards, are then required to allow the user to take an important number of pictures, stored into a simple memory card. In brief, the purpose of the coding step is to reduce the redundancy in the image and to remove insignificant content to match the capacity of the storage device. Compression algorithms require however the whole image for, finally, discarding an important part of (irrelevant) information. This may appear to be wasteful for applications whose sampling scheme is expensive to perform. Many computing resources could then be saved up if the compressed coefficients were directly acquired out of the sensor.

Recently, a new theory of sampling has been emerged in the signal processing community. This theory, introduced as the Compressive Sampling or Compressed

Sensing [Candès 2006b, Donoho 2006], suggests that one can reconstruct perfectly a signal, supposed to be sparse in some basis, from a limited (i.e. fewer than Nyquist) number of incoherent measurements. The motivation behind the CS technique is to perform in the same time the acquisition and the compression of the signal. We give a quick overview of this technique in this section but more information can be found in the referred works.

6.1.2 Main results

Let $x_0 \in \mathbb{R}^N$ be a Nyquist sampled version of the analog measured scene. The main result of the CS theory states that x_0 can be recovered exactly from a small number of measurements [Candès 2006a] directly outcomed from the sensor. The key of the CS theory relies on the supposed sparsity of the original signal x_0, meaning that it can be perfectly represented in some basis $\Psi : \mathbb{R}^N \rightarrow \mathbb{R}^N$ with only S non-null coefficients. This property of sparsity is actually well-known for natural images and widely used by coding algorithms to represent the content of images on compact bitstreams [Wallace 1992].

Based on this property of sparsity, the authors of [Candès 2006a] showed that only M (with $M << N$) measurements are required to perfectly reconstruct the original signal x_0 with a high probability. These M observations are obtained by the projection of the image $x_0 \in \mathbb{R}^N$ on a measurement matrix $\Phi : \mathbb{R}^N \rightarrow \mathbb{R}^M$

$$y = \Phi x_0. \tag{6.1}$$

Matrix Φ being not of full rank, it seems difficult to recover x_0 exactly. However, it appears that if one considers the image x_0 to be sparse in some basis Ψ, then all the information and the structure of x_0 is conserved in y with a high probability [Candès 2006a]. More precisely, let $\alpha_0 \in \mathbb{R}^N$ be a S sparse vector, that is a vector having S non-null coefficients, and let $y \in \mathbb{R}^M$ be the measurement vector obtained by

$$y = \Phi \alpha_0. \tag{6.2}$$

If we assume that we know the location of the S sparse coefficients, only S linearly independent equations are then required to recover α_0 from y. In otherwords, one can recover α_0 exactly from y if the sub-matrix Φ_K of size $M \times S$ is full rank. The restricted isometry property (RIP) has been introduced in [Candès 2006d, Candès 2006a] to generalize this notion of quasi-orthonormality. Let $\theta \in \mathbb{R}^N$ be a S sparse vector, then the measurement matrix Φ owns the RIP of order S if for any sub-matrix Φ_p of size $M \times p$ with $p \in [1, \ldots, S]$, one has

$$(1 - \delta_K)\|\theta\|_2^2 \leq \|\Phi_p \theta\|_2^2 \leq (1 + \delta_K)\|\theta\|_2^2, \tag{6.3}$$

where δ_K is the smallest constant (known as the restricted isometry constant) which verifies (6.3) for any p. The design of such measurements matrices is however a NP-complete task. Fortunately, it appears that most of random matrices, such as Gaussian random matrices or matrices outcomed from Bernouilli processes

[Candès 2006a], satisfy the RIP of order $2K$ (required to preserve the distance between any two sparse signals) with a high probability [Baraniuk 2008].

When no prior information on the location of the non-null sparse coefficients is available, recovering α_0 from y is more difficult. The authors of [Donoho 2006] addressed this problem and showed that, if the RIP condition is satisfied, the image x_0 can be recovered with a high probability by minimizing the l^0-norm of its coefficients in Ψ, under the constraint that its projection on Φ is equal to the observed vector y. This however leads to a NP-complete algorithm [Donoho 2006]. A strong result, due to [Candès 2006b], states that the l^0-norm can be equivalently replaced by the l^1-norm. The reconstruction problem is then formulated as follows [Candès 2006b]

$$\text{Find } \tilde{x} \in \begin{array}{ll} \arg\min & \|\Psi x\|_1 \\ \text{subject to} & x \in \mathbb{R}^N \\ & y = \Phi x \end{array} . \tag{6.4}$$

The optimization problem (6.4) is a particular instance of the Basis Pursuit problem [Chen 1998] which can be efficiently solved using classical algorithms from the linear programming literature. Problem (6.4) can be interpreted as follows. The randomness of the measurement matrix Φ spreads the content of the image in the measurement vector y. If Φ satisfies the RIP, then the inverse solution $\Phi^\dagger y$ contains all the information of the image x_0 but in disorder. Also remind that the representation of x_0 in the basis Ψ is sparse or, in other words, strongly compact. Minimizing the l^1-norm of its coefficients will then put the non-null coefficients back at the correct position, recovering therefore the original image.

It is shown in [Candès 2007] that solving problem (6.4) leads to an exact solution if x_0 is sparse enough in Ψ. Therefore, the more sparse is x_0 the easier it will be for the algorithm (6.4) to recover the original signal. Recovering the image x_0 highly depends on the link between the compactness of the decomposition basis Ψ and the diffusion of the measurement matrix Φ. More generally, the algorithm (6.4) efficiently recovers the original image only if matrices Φ and Ψ are completely uncorrelated. A mutual coherence μ has been introduced in [Candès 2007] to measure this correlation and more precisely, to measure the correlation between each vector basis ϕ_i and ψ_j of Φ and Ψ. It is defined as

$$\mu(\Phi, \Psi) = \sqrt{N} \max_{i,j} |\langle \phi_i, \psi_j \rangle|. \tag{6.5}$$

This coherence measure belongs to $[1, \sqrt{N}]$ [Candes 2008]; a small value of μ meaning that the matrices Ψ and Φ are completely uncorrelated. For example, if Φ is the Fourier basis, then the minimal coherence is obtained with $\Psi = I$ (the sampling operator) and is equal to 1. Such scenario actually corresponds to magnetic resonance imaging for example, where the data is directly acquired in the Fourier domain [Lustig 2007]. More generally, solving (6.4) recovers x_0 exactly if

$$M \geq C\mu^2(\Phi, \Psi) S \log(N), \quad C < 1 \text{ is a constant.} \tag{6.6}$$

In classical imaging systems, acquired images are usually degraded by both blur and instrumental noise. As shown in [Jianwei 2009], the CS technique is robust to this scheme. In the case of blurred and noisy measurements, the acquisition model (6.1) becomes [Jianwei 2009]

$$y = \Phi H x_0 + z, \tag{6.7}$$

where $H : \mathbb{R}^N \to \mathbb{R}^N$ is the blur matrix and $n \in \mathbb{R}^M$ is an additive noise. In the classical case of an additive white Gaussian noise of variance σ_n^2, the reconstruction algorithm may write [Jianwei 2009]

$$\text{Find } \tilde{x} \in \begin{array}{rl} \arg\min & \|\Psi x\|_1 \\ \text{subject to} & x \in \mathbb{R}^N \\ & \|y - \Phi H x\|_2^2 \leq M\sigma_z^2 \end{array}. \tag{6.8}$$

Similarly to (6.4), the optimization problem (6.8) is a particular instance of Basis Pursuit Denoising which can also be solved using linear programming techniques [Chen 1998]. Of course, exact reconstruction cannot be achieved anymore due to the error on the measurements introduced by the noise. The reconstruction error can however be accurately estimated (at least in the case of measurements only degraded by noise) as a function of the restricted isometry constant [Candès 2006c].

Although the design of a sensor able to produce these random measurements is difficult and beyond the scope of the thesis, the CS technique clearly appears to be adapted to the satellite imaging chain. It could indeed drastically simplify the process of image acquisition by providing a reduced number of measurements, directly outcomed from the sensor, therefore saving an important quantity of resources. It is also valuable to point out that the CS framework provides an acquisition technique whose performances depend mainly on the reconstruction algorithm done on-ground. In comparison, the current acquisition imaging chain is bounded by the efficiency of the compression scheme embedded on-board. In that case, if one wants to increase the quality of the final image, one has to design a new image coder. This "universal" coding feature [Candès 2006d] of the CS is thus very attractive. In the next part, we propose therefore a satellite imaging chain based on this technique. We formulate the acquisition model and we present an algorithm to reconstruct the image from the measurements vector.

6.2 Compressed Sensing based satellite imaging chain

6.2.1 Acquistion model of the satellite imaging chain

As said previously, we assume that we have at our disposal a sensor able to produce incoherent measurements, in the sense of the CS framework. We are interested in evaluating the quality of the reconstructed image in comparison to the image obtained using the current acquisition chain based on wavelet compression [Antonini 1992].

Though it is not a general result (see [Goyal 2008] for example), previous works [Schulz 2009] have shown that the CS technique may be competitive regarding to

a wavelet-based compression scheme on smoothed classical test images. But to the best of our knowledge, no works have been dedicated to this comparison for high-resolution satellite imaging, taking into account the degradations of the satellite imaging acquisition chain (blur, instrumental and quantizing noises).

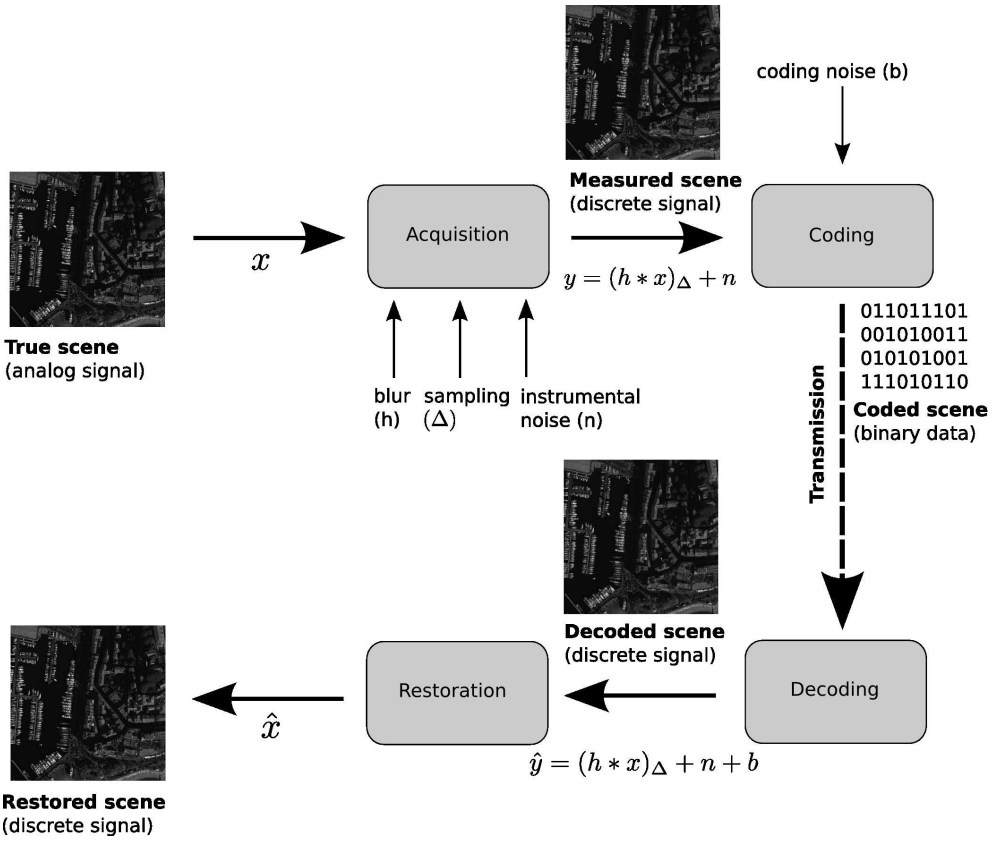


Figure 6.1: Current satellite imaging chain.

The current satellite imaging chain used by the CNES is recalled Fig. 6.1. In the case of a CS based acquisition technique, the instrumental image at the output of the acquisition can be written as the projection of the blurred image on the measurement matrix Φ, noised by an instrumental noise n

$$y = \Phi H x + n, \tag{6.9}$$

where H is the matrix notation for the PSF, n is the instrumental noise supposed to be a zero-mean Gaussian distribution with a known variance σ^2. We assume that the variance of this noise is pixel dependent and we use the model (3.3) to express this dependence.

In addition to blur and instrumental noise, the M measurements are also degraded by quantizing noise. In a classical satellite imaging chain, a wavelet transform is usually applied prior quantizing to decorrelate the data. Since, the acquired data is random, in the CS technique, and does not present any favored structure, we propose here to directly quantize the coefficients y. We modelize this quantization Q as

a scalar uniform quantization which quantizing step Δ_i depends on the coefficient $y_i, i \in \{1, \dots, M\}$ regarded

$$Q(y_i) = \Delta_i \left\lfloor \frac{y_i}{\Delta_i} + \frac{1}{2} \right\rfloor, \tag{6.10}$$

where $\lfloor . \rfloor$ is the floor function which returns the greatest integer less than or equal to its argument. The quantizing step Δ_i can be transmitted with the image as in the JPEG standard [Wallace 1992] or can be deduced during the decoding algorithm for more recent methods [Taubman 2000, Said 1996]. Therefore, we assume in the following that the quantizing steps Δ_i are known. Let $b = Q(y) - y$ be the quantizing error. From (6.10), we have for each coordinate b_i of b

$$-\frac{\Delta_i}{2} \leq b_i < \frac{\Delta_i}{2}, \quad \forall i \in \{1, \dots, M\} \tag{6.11}$$

or equivalently

$$b \in B, \quad \text{with } B = \left\{ b \in \mathbb{R}^M, -\frac{\Delta_i}{2} \leq b_i < \frac{\Delta_i}{2} \quad \forall i \in \{1, \dots, M\} \right\}. \tag{6.12}$$

Using the previous definition of b, we propose to modelize the observed measurements as

$$\hat{y} = Q(\Phi H x + n) = \Phi H x + n + b, \tag{6.13}$$

where $\hat{y}$ is the measurements vector.

6.2.2 Proposed reconstruction algorithm

The extension of the reconstruction algorithm (6.8) to the acquistion model (6.13) is simple. First, simply remark that the problem (6.8) can also be written

$$\text{Find } \tilde{x} \in \begin{array}{ll} \arg\min & \|\Psi x\|_1 \\ \text{subject to} & x \in \mathbb{R}^N, n \in \mathbb{R}^M \\ & \|\frac{1}{\sigma_z} n\|_2^2 \leq M \\ & y = \Phi H x + n \end{array}. \tag{6.14}$$

In our case, the variable b needs to be added to the problem (6.14) to take into account the presence of the coding noise. Using (6.12) and (6.13), the reconstruction problem for acquisition model (6.13) writes

$$\text{Find } \tilde{x} \in \begin{array}{ll} \arg\min & \|\Psi x\|_1 \\ \text{subject to} & x \in \mathbb{R}^N,\ n \in \mathbb{R}^M,\ b \in \mathbb{R}^M \\ & \|\Sigma n\|_2^2 \leq M, \\ & b \in B, \\ & \hat{y} = \Phi H x + n + b \end{array}. \tag{6.15}$$

where $\Sigma = \text{diag}\left(\frac{1}{\sigma_i}\right)$ is used to take into account the pixel dependence of the variance of the noise n. The problem (6.15) can be further simplified by noting that

the variable b can be replaced by $\hat{y} - (\Phi H x + n)$. We finally propose to formulate the reconstruction problem as

$$\text{Find } \tilde{x} \in \begin{array}{ll} \arg\min & \|\Psi x\|_1 \\ \text{subject to} & x \in \mathbb{R}^N,\ n \in \mathbb{R}^M \\ & \|\Sigma n\|_2^2 \leq M, \\ & \hat{y} - (\Phi H x + n) \in B \end{array} . \tag{6.16}$$

The optimization problem (6.16) is a convex problem constrained on convex sets and thus admits a unique (convex) set of solutions [Boyd 2004]. However, the presence of the linear operators Ψ, Φ and H make it difficult to solve.

We propose here to use the alternating direction method of multipliers proposed in [Afonso 2011]. The advantage of this algorithm is that it is very general and it gives satisfying computing time. It solves

$$\text{Find } (\tilde{u}, \tilde{v}) \in \begin{array}{ll} \arg\min & f_1(u) + f_2(v) \\ \text{subject to} & Cu + Dv = a \\ & u \in \mathbb{R}^p, v \in \mathbb{R}^q \end{array} , \tag{6.17}$$

where

- $f_1 : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ and $f_2 : \mathbb{R}^q \to \mathbb{R} \cup \{+\infty\}$ are two closed convex functions.
- $C \in \mathbb{R}^{l \times p}$ and $D \in \mathbb{R}^{l \times q}$ are two linear operators.
- $a \in \mathbb{R}^l$ is a given vector.

The alternating direction algorithm relies on the augmented Lagrangian method. Let $\lambda \in \mathbb{R}^l$ be a Lagrange multiplier attached to the linear constraint (6.17), the augmented Lagrangian writes

$$\mathcal{L}(u, v, \lambda) = f_1(u) + f_2(v) + \langle \lambda, Cu + Dv - a \rangle + \frac{\beta}{2} \|Cu + Dv - a\|_2^2, \tag{6.18}$$

where β is a parameter which controls the linear constraint [Glowinski 1984]. This parameter has to belong to the interval $\left]0, \frac{\sqrt{5}+1}{2}\right[$ to ensure that $\{(u^k, v^k)\}$ converge to the set of minimizers [Glowinski 1984].

This algorithm consists in finding a saddle point of the augmented Lagrangian, thereby solving (6.17), by minimizing it in an alternating way, subject to u, v, then to λ. The algorithm is given in algorithm 4.

We now detail how to apply algorithm 4 to problem (6.16). To match the class of problem (6.17), we define

$$u = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} \in R^N \times R^M \times R^M, v = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} x \\ n \end{pmatrix} \in R^M \times R^M,$$

$$a = \begin{pmatrix} 0 \\ 0 \\ -\hat{y} \end{pmatrix} \in R^N \times R^M \times R^M, \tag{6.19}$$

Algorithm 4 Alternating direction method of multipliers to solve (6.17)

Set the number of iterations K.
Set an initial point $u^0 \in \mathbb{R}^p$.
Set an initial point $v^0 \in \mathbb{R}^q$.
Set an initial point $\lambda^0 \in \mathbb{R}^l$.
Set $\gamma > 0$ and $\beta > 0$.
for k from 0 to $K-1$ **do**
Compute $u^{k+1} = \begin{array}{rl} \arg\min & \mathcal{L}(u, v^k, \lambda^k) \\ \text{subject to} & u \in \mathbb{R}^p \end{array}$.
Compute $v^{k+1} = \begin{array}{rl} \arg\min & \mathcal{L}(u^{k+1}, v, \lambda^k) \\ \text{subject to} & v \in \mathbb{R}^q \end{array}$.
Set $\lambda^{k+1} = \lambda^k + \beta\gamma(Cu^{k+1} + Dv^{k+1} - a)$.
end for

and

$$C = I, \tag{6.20}$$

$$D = \begin{bmatrix} \Psi & 0 \\ 0 & I \\ -\Phi H & -I \end{bmatrix}, \tag{6.21}$$

where I is the identity matrix. Using these definitions, problem (6.16) can be reformulated

$$\text{Find } (\tilde{u}, \tilde{v}) \in \begin{array}{rl} \arg\min & \|u_1\|_1 \\ \text{subject to} & u \in \mathbb{R}^N \times \mathbb{R}^M \times \mathbb{R}^M,\ v \in \mathbb{R}^M \times \mathbb{R}^M \\ & \|\Sigma u_2\|_2^2 \leq M, \\ & u_3 \in B, \\ & -u + Dv = a \end{array}. \tag{6.22}$$

We further define

$$f_2(v) = 0, \tag{6.23}$$

$$f_1(u) = \|u_1\|_1 + \chi_G(u_2) + \chi_B(u_3), \tag{6.24}$$

where χ_G is the indicator function on a weighted l^2 ball

$$\chi_G(u_2) = \begin{cases} 0, & \text{if } \|\Sigma u_2\|_2^2 \leq M \\ \infty, & \text{otherwise} \end{cases}, \tag{6.25}$$

and χ_B is the indicator function on the hypercube B

$$\chi_B(u_3) = \begin{cases} 0, & \text{if } u_3 \in B \\ \infty, & \text{otherwise} \end{cases}. \tag{6.26}$$

Using these notations, it is straightforward to see that problem (6.16) fits the formulation (6.17) and becomes

$$\text{Find } (\tilde{u},\tilde{v}) \in \begin{array}{rl} \arg\min & f_1(u) \\ \text{subject to} & u \in \mathbb{R}^N \times \mathbb{R}^M \times \mathbb{R}^M,\ v \in \mathbb{R}^M \times \mathbb{R}^M \\ & -u + Dv = a \end{array}. \tag{6.27}$$

The first step of the algorithm consists in computing

$$u^{k+1} = \begin{array}{rl} \arg\min & \mathcal{L}(u, v^k, \lambda^k) \\ \text{subject to} & u \in R^N \times \mathbb{R}^M \times \mathbb{R}^M \end{array}, \tag{6.28}$$

where $\mathcal{L}$ is the augmented Lagragian which, for problem (6.27), writes

$$\mathcal{L}(u,v,\lambda) = f_1(u) + \langle \lambda, Dv - u - a\rangle + \frac{\beta}{2}\|Dv - u - a\|_2^2. \tag{6.29}$$

We have

$$\begin{aligned} u^{k+1} = & \begin{array}{rl} \arg\min & f_1(u) + \langle \lambda, Dv^k - u - a\rangle + \frac{\beta}{2}\|Dv^k - u - a\|_2^2 \\ \text{subject to} & u \in \mathbb{R}^N \times \mathbb{R}^M \times \mathbb{R}^M \end{array} \\ = & \begin{array}{rl} \arg\min & \frac{1}{\beta} f_1(u) + \frac{1}{2}\|Dv^k - a + \frac{\lambda^k}{\beta} - u\|_2^2 \\ \text{subject to} & u \in \mathbb{R}^N \times \mathbb{R}^M \times \mathbb{R}^M \end{array} \\ = & \operatorname{prox}_{\frac{1}{\beta} f_1}\left(Dv^k - a + \frac{\lambda^k}{\beta} \right), \end{aligned} \tag{6.30}$$

where prox is the proximal operator presented in [Combettes 2005]. For any function $f : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$, the proximal operator prox_f is defined by

$$\operatorname{prox}_f\left(x_0 \right) = \underset{x \in R^N}{\arg\min}\, f(x) + \frac{1}{2}\|x - x_0\|_2^2. \tag{6.31}$$

We recall two results of [Combettes 2005] that we will use. Let $X \subseteq \mathbb{R}^N$ be a closed convex set and $f(x) = \begin{cases} 0 & \text{if } x \in X \\ +\infty & \text{otherwise} \end{cases}$. Then

$$\operatorname{prox}_f = \Pi_X, \tag{6.32}$$

where Π_X is the euclidian projector on the set X. It is straightforward to see that the proximal operator generalizes the notion of projection. If $f(x) = \tau\|x\|_1$, then prox_f is the soft-thresholding operator and we have

$$\operatorname{prox}_{\tau\|\cdot\|_1}\left(x_0 \right) = \operatorname{shrink}_\tau(x_0) = \operatorname{sign}(x_0)\max(|x_0| - \tau, 0). \tag{6.33}$$

Using results (6.32) and (6.33), we have

$$u^{k+1} = \begin{pmatrix} \operatorname{shrink}_{\frac{1}{\beta}}\left(Dv^k - a + \frac{\lambda^k}{\beta}\right) \\ \Pi_G\left(Dv^k - a + \frac{\lambda^k}{\beta}\right) \\ \Pi_B\left(Dv^k - a + \frac{\lambda^k}{\beta}\right) \end{pmatrix}, \tag{6.34}$$

where Π_G is the orthogonal projection on a weighted l^2 ball and Π_B is the orthogonal projection on the hypercube B. The projection Π_B is simple to compute and writes

$$(\Pi_B(x_0))_i = \begin{cases} (x_0)_i & \text{if } -\frac{\Delta_i}{2} \leq (x_0)_i < \frac{\Delta_i}{2} \\ -\frac{\Delta_i}{2} & \text{if } (x_0)_i < -\frac{\Delta_i}{2} \\ \frac{\Delta_i}{2} & \text{if } \frac{\Delta_i}{2} \leq (x_0)_i \end{cases} . \tag{6.35}$$

The projection Π_G is more difficult to address and can be solved efficiently using an iterative scheme. This projection is detailled in [Weiss 2009] and we refer the interested reader to this paper for the computation of this projection.

The second step of the algorithm requires to compute v^{k+1}. We have

$$\begin{aligned} v^{k+1} = & \quad \arg\min \quad \langle \lambda^k, Dv - u^k - a \rangle + \frac{\beta}{2}\|Dv - a - u^k\|_2^2 \\ & \text{subject to} \quad v \in \mathbb{R}^M \times \mathbb{R}^M \\ = & \quad \arg\min \quad \frac{\beta}{2}\|Dv + \frac{\lambda^k}{\beta} - a - u^k\|_2^2 \\ & \text{subject to} \quad v \in \mathbb{R}^M \times \mathbb{R}^M \end{aligned} . \tag{6.36}$$

v^{k+1} is then the solution of the positive-semidefinite linear system

$$D^*Dv = D^*\left(a + u^k - \frac{\lambda^k}{\beta}\right). \tag{6.37}$$

Equation (6.37) needs then to invert D^*D. Most of the time, the operator D^*D owns a particular structure which can be numerically exploited to solve (6.37). This remark has been used in [Ng 2010] for example to obtain fast algorithms. System (6.37) can also be solved using standard techniques such as conjugate gradient. In our experiments we observe that 10 iterations of a conjugate gradient method are sufficient to solve (6.37). Note that sub-problems (6.28) and (6.36) can be solved approximately while preserving the convergence of the algorithm [He 2002].

The resulting algorithm is given in the algorithm 5.

Algorithm 5 Alternating direction method of multipliers to solve (6.16)

Set the number of iterations K.
Set an initial point $u^0 \in \mathbb{R}^p$.
Set an initial point $v^0 \in \mathbb{R}^q$.
Set an initial point $\lambda^0 \in \mathbb{R}^l$.
Set $\gamma > 0$ and $\beta > 0$.
for k from 0 to $K-1$ **do**
  Compute u^{k+1} from (6.34).
  Compute v^{k+1} by solving (6.37).
  Set $\lambda^{k+1} = \lambda^k + \beta\gamma(Dv^{k+1} + u^{k+1} - a)$.
end for
Output $\tilde{x} = v_1$.

6.2.3 Numerical results

We evaluate the performances of the CS technique for satellite imaging in comparison to the chain used by the CNES and based on a wavelet compression scheme. For the numerical experiments, we choose the measurement matrix Φ to be the noiselet transform [Coifman 2001] and set Ψ to be the gradient operator such that $||\Psi x||_1$ is the TV [Rudin 1992]. We made the choice of the TV as it is almost equivalent to a Haar basis which, as required by the CS framework, shares a small mutual coherence with the noiselet transform [Candes 2008].

As mentioned previously, we compare the CS acquisition technique to the classical acquisition chain which consists in sampling the real image at the Nyquist frequency followed by a compression scheme. The considered compression algorithm uses the biorthogonal CDF 9/7 wavelet transform described in [Cohen 1992] followed by the same quantization process as the one defined in (6.10). In that case, the aquisition model writes

$$\hat{y} = Q(W(Hx + n)), \tag{6.38}$$

where W is the CDF 9/7 wavelet transform. As in the CS technique, we can design an algorithm to reconstruct the image from the noisy observed wavelet coefficients $\hat{y}$

$$\text{Find } \tilde{x} \in \begin{array}{ll} \arg\min & \|\Psi x\|_1 \\ \text{subject to} & x \in \mathbb{R}^N,\ n \in \mathbb{R}^N \\ & \hat{y} - W\left(Hx + n\right) \in B \\ & \|\Sigma n\|_2^2 \leq N \end{array}, \tag{6.39}$$

where Ψ is the gradient operator. Note that the formulation (6.39) is not expressed using any matrix Φ as, in this case, the measurement matrix is the sampling operator ($\Phi = I$). We will compare the results of techniques (6.16) and (6.39) visually but also in a rate-distortion sense. As both techniques offer different ways to control the target coding rate, we now detail the choice of the coding parameters in each case.

For the CS technique, we take benefit from the fact that the image can ideally be reconstructed from less measurements than Nyquist. More precisely, for a low target rate, we will restrict the number of measurements M to be small and when the target rate is high, we will increase this number, the maximum number of measurements being equal to the number of pixels N. This particular choice comes from the fact that the distribution of the CS coefficients is quite large and that a high quantization has to be applied on these coefficients to reach low target rates [Fletcher 2007]. It seems then more appropriate to tune the number of measurements M instead of tuning the quantizing steps, for a given coding rate. Consequently, we will always take $\Delta_i = 1, \forall i \in \{1, \ldots, M\}$ for all coding rates. Note that these measurements will be taken randomly and that the position of the retained coefficients can be known at each side of the chain by transmitting the seed of the random generator.

The imaging chain based on a wavelet scheme does not however offer such feature. More precisely, all the coefficients have to be retained to be able to reconstruct the

image. Consequently, for this technique, we will keep all the coefficients and we will tune the quantizing steps to reach the target coding rate. For more simplicity, we will take the same quantizing step for all coefficients $\Delta_i = \Delta, \forall i \in \{1, \dots, N\}$.

As mentioned previously, we will evaluate the results in a rate-distortion sense. The distortion will be evaluated using the PSNR defined in (4.98). For the evaluation of the coding rate, we assume that the quantized coefficients will be encoded using an entropy encoder. The coding rate R can then be measured using the entropy (expressed in bits/symbol) of the coefficients $\hat{y}$ [Shannon 1948]

$$R(\hat{y}) = -\sum_{m=-\infty}^{\infty} p_{\hat{y}}(m) \log_2(p_{\hat{y}}(m)), \tag{6.40}$$

where $p_{\hat{y}}(m)$ is the probability for a quantized coefficient to get the symbol m. Note that for the imaging chain based on the CS technique, we only retain M coefficients. Since the sampling of these M is done randomly, one can transmit the seed of the random generator to reproduce the same sampling scheme. The position of the M coefficients can thus be assumed to be known by the decoder, without the need to transmit more information than a seed (which holds on a few bytes), and does not have to be taken into account in the computation of the entropy. The entropy of the quantized coefficients will be thus multiplied by the ratio between the number of measurements and the number of pixels for that case.

We simulate the two imaging chains on the reference image depicted Fig. 6.2. The blur H used in this simulation is the PSF provided by the CNES and the instrumental noise n is a zero-mean Gaussian noise with variance given by (3.3).

Results are shown on Fig. 6.3 and 6.4. From the rate-distortion function displayed on Fig. 6.3, we see that the CS technique does not give competitive reconstruction results in comparison to the wavelet-based technique, and stands $5-6$ dB below this technique, for all compression rates. Visually, the reconstructed images are not very good as well. We can see on Fig. 6.4 that the CS reconstruction algorithm overregularizes the solution and creates large patterns, therefore losing the details of the image. Although it seems clear that the CS is a good acquisition technique as it better spreads the information than a wavelet transform, it also appears that high-resolution satellite images are not sparse enough, in usual basis, such that this technique is difficult to apply.

Moreover, as said previously the CS coefficients have a large distribution (larger than wavelet coefficients) making their coding difficult to perform, even when one only retains a limited number of these coefficients. We have however strong thoughts that the CS could be an efficient acquisition strategy for satellite images as it has already shown interesting results in application where the image is naturally strongly sparse, such as in MRI application [Lustig 2007]. Following this idea, an imaging chain based on the CS technique may be interesting for galaxy observation missions which naturally give sparse images, as in astronomy where the CS exhibits great performances [Bobin 2008]. Due to time constraint, this aspect has not been addressed in the thesis.

Figure 6.2: Reference image, Cannes harbour (12 bits panchromatic image, 30 cm resolution, 1024×1024 pixels).

However, in the case of earth observation missions, the approximative sparsity of satellite images does not seem to be sufficient to make the CS technique competitive regarding to the classical wavelet approach.

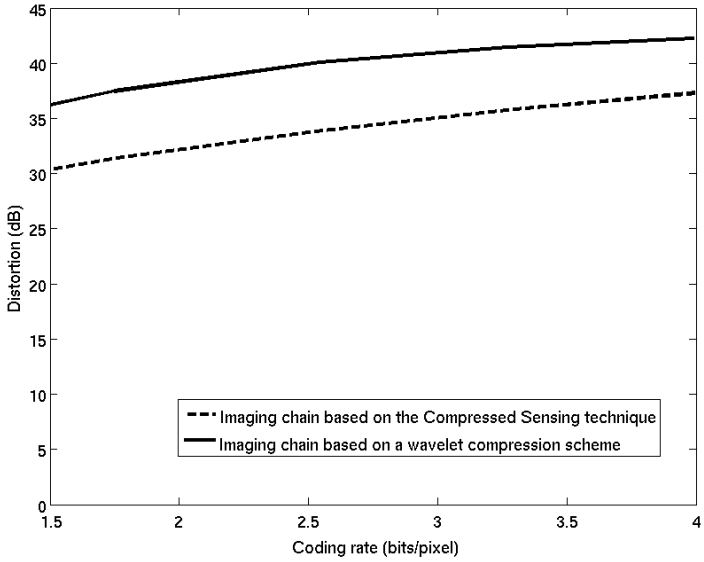


Figure 6.3: Rate-distortion function for the two acquisition techniques. The dashed curve is the PSNR w.r.t. the compression rate for the CS acquisition technique while the solid curve is the PSNR w.r.t. the compression rate for the wavelet-based method.

6.3 Conclusion and perspectives

In this part, we have experimentaly studied the performances of the CS acquisition technique in application to satellite imaging. We showed that this technique is interesting for satellite imaging chain since it proposes a low-resources acquisition technique which matches the reduced embedded computational capacity of satellites.

We proposed a novel imaging chain based on this framework and we formulated a decoding algorithm which takes into account the main degradations of the satellite imaging chain (blur, instrumental and quantizing noise). We showed reconstruction results, visually and in a rate-distortion sense, on a real satellite data and we performed a comparison of this method to the classical acquisition method based on a wavelet transform.

The obtained results showed that the CS acquisition technique does not give competitive results for earth observation imaging since satellite images of such application can not be represented in a compact form using classical transform, i.e. the information of the image can not be contained on a reduced number of coefficients. The CS acquisition method could be however interesting for galaxy observation

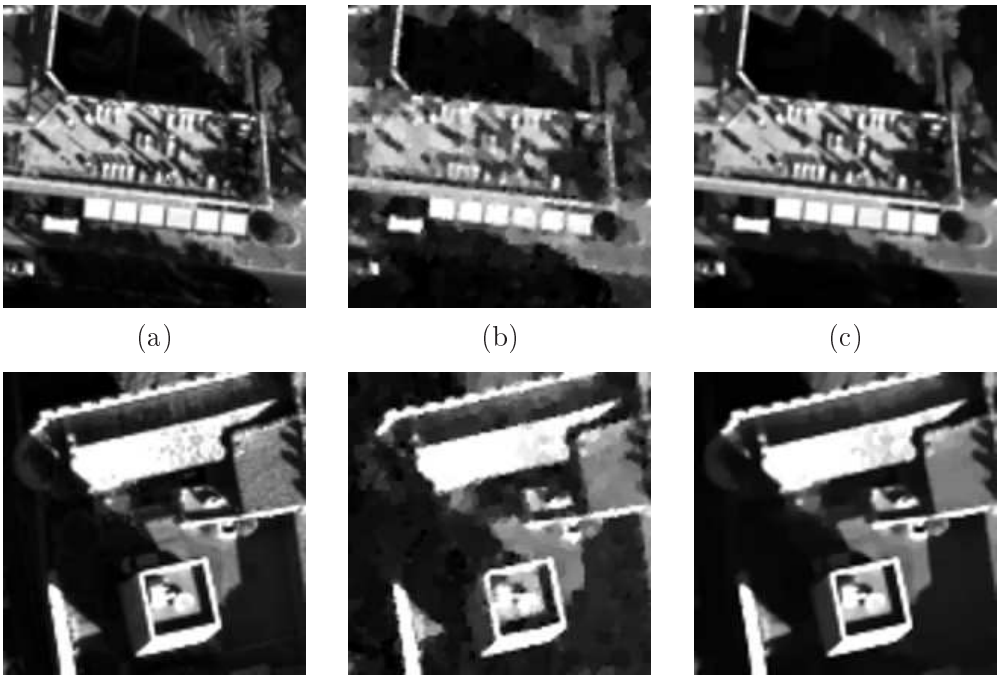


Figure 6.4: Reconstruction results for the two acquisition techniques at a compression rate of 2.5 bits/pixel. (a)-(d) are zooms of the original image, (b)-(e) are zooms of the reconstructed image using the CS technique ($PSNR = 33.8\ dB$) and (c)-(f) are zooms of the reconstructed image using the wavelet-based technique ($PSNR = 40\ dB$).

missions which give images which are naturally compact. There is also room for improvements by considering more properly the distribution of satellite images to enhance usual priors, used by the decoder, and quantizing strategies which need to better fit the characteristics of the CS coefficients distribution.

Part IV

Conclusion

Chapter 7

Conclusion of the thesis

This chapter is the conclusion of the thesis. It summarizes the contributions of the thesis and discusses some perspectives of this work.

7.1 Conclusion and summary of the contributions

In this thesis we addressed the problem of imaging chain optimization in the context of satellite imaging and we proposed several methods which focus on the problem of global optimization of the compression/restoration chain.

Formulating an expression of the global distortion is a difficult task since many intermediate variables are correlated. In this thesis, we presented a method to solve this problem and we achieved to theoretically estimate the global distortion of a simple case of imaging chain. We then proposed an algorithm to minimize the estimated distortion with respect to the parameters of the chain. We also developed the proposed method for three different configuration of the imaging chain to address the question of the optimal position of the restoration in the imaging chain.

We also presented, in the thesis, an alternative method to optimize of the quality of the final image. Though this study is mainly experimental, we succeeded to address recurrent open questions such as the position of the restoration in the chain and how to deal with the coding noise. From the obtained results, we proposed a new satellite imaging chain which eliminates several current problems in the observation of the final image.

Finally, we presented in the last part of the thesis a novel satellite imaging chain based on a recent theory of sampling. We showed that low-resources sampling technique is interesting for satellite imaging and we proposed an algorithm to solve the reconstruction problem.

7.2 Perspectives

Several future investigations may be opened to improve the results obtained in this thesis.

The extension of the imaging chain that we considered in Chapter 4 to the true satellite seems difficult to achieve. An alternative technique to express the global distortion as a function of the chain parameters may then consists in using the unbiased estimators presented in [Ramani 2008]. In that case, the difficulty is to extend these estimators to the acquisition model of the satellite imaging chain which is complex.

Regarding to the imaging chain that we proposed in Chapter 5, it would be very interesting to study how to limit the power of the residual noise on the final image. Since this residual noise depends on the target coding rate, it may be interesting to focus on advances coding techniques with the challenge to conserve the decorrelation property of dithering techniques. Conversely, it would be worth extending the actual subtractive dithering techniques, used by the proposed imaging chain, to match more complex quantizing schemes, similarly to [Stamm 2011].

Finally, an interesting investigation for the CS acquisition technique would be to evaluate its performances on naturally sparse satellite images like the ones obtained from galaxy observation missions.

Bibliography

[Afonso 2011] M.V. Afonso, J.M. Bioucas-Dias and M.A.T. Figueiredo. *An Augmented Lagrangian Approach to the Constrained Optimization Formulation of Imaging Inverse Problems.* IEEE Transactions on Image Processing, vol. 20, no. 3, pages 681–695, Mar. 2011. (Cited on page 133.)

[Andrews 1971] H.C. Andrews. *Multidimensional Rotations in Feature Selection.* IEEE Transactions on Computers, vol. 20, pages 1045–1051, 1971. (Cited on page 16.)

[Antonini 1992] M. Antonini, M. Barlaud, P. Mathieu and I. Daubechies. *Image coding using wavelet transform.* IEEE Transactions on Image Processing, vol. 1, no. 2, pages 205–220, Apr. 1992. (Cited on pages 18, 25, 51, 90 and 130.)

[Baraniuk 2008] R. Baraniuk, M. Davenport, R. Devore and M. Wakin. *A simple proof of the restricted isometry property for random matrices.* Constructive Approximation, vol. 23, pages 253–263, 2008. (Cited on page 129.)

[Bect 2004] J. Bect, L. Blanc-Féraud, G. Aubert and A. Chambolle. *A l1-unified variational framework for image restoration.* In Proceedings of European Conference on Computer Vision, pages 1–13, 2004. (Cited on pages 24, 25 and 93.)

[Berger 1971] T. Berger. Rate distortion theory: A mathematical basis for data compression. Prentice-Hall, 1971. (Cited on page 90.)

[Bobin 2008] J. Bobin, J.-L. Starck and R. Ottensamer. *Compressed Sensing in Astronomy.* IEEE Journal of Selected Topics in Signal Processing, vol. 2, no. 5, pages 718–726, Oct. 2008. (Cited on page 138.)

[Boyd 2004] S. Boyd and L. Vandenberghe. Convex optimization. Cambridge University Press, 2004. (Cited on pages 55 and 133.)

[Candès 2005] E.J. Candès and T. Tao. *Decoding by linear programming.* IEEE Transactions on Information Theory, vol. 51, no. 12, pages 4203–4215, Dec. 2005. (Cited on page 27.)

[Candès 2006a] E.J. Candès, L. Demanet, D.L. Donoho and L. Ying. *Fast discrete curvelet transforms.* Multiscale Modeling and Simulation, vol. 5, no. 3, pages 861–899, 2006. (Cited on pages 93, 128 and 129.)

[Candès 2006b] E.J. Candès, J. Romberg and T. Tao. *Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information.* IEEE Transactions on Information Theory, vol. 52, no. 2, pages 489–509, Feb. 2006. (Cited on pages 128 and 129.)

[Candès 2006c] E.J. Candès, J.K. Romberg and T. Tao. *Stable signal recovery from incomplete and inaccurate measurements.* Communications on Pure and Applied Mathematics, vol. 59, no. 8, pages 1207–1223, Aug. 2006. (Cited on page 130.)

[Candès 2006d] E.J. Candès and T. Tao. *Near-Optimal Signal Recovery From Random Projections: Universal Encoding Strategies?* IEEE Transactions on Information Theory, vol. 52, no. 12, pages 5406–5425, Dec. 2006. (Cited on pages 128 and 130.)

[Candès 2007] E.J. Candès and J.K. Romberg. *Sparsity and incoherence in compressive sampling.* Inverse Problems, vol. 23, no. 3, pages 969–985, 2007. (Cited on page 129.)

[Candes 2008] E.J. Candes and M.B. Wakin. *An Introduction To Compressive Sampling.* IEEE Signal Processing Magazine, vol. 25, no. 2, pages 21–30, Mar. 2008. (Cited on pages 129 and 137.)

[CCSDS 2005] CCSDS. *Image Data Compression.* Blue Book, Nov. 2005. (Cited on pages 37, 39 and 91.)

[Chambolle 1997] A. Chambolle and P.-L. Lions. *Image recovery via total variation minimization and related problems.* Numerische Mathematik, vol. 76, no. 2, pages 167–188, Apr. 1997. (Cited on page 25.)

[Chappelier 2006] V. Chappelier and C. Guillemot. *Oriented Wavelet Transform for Image Compression and Denoising.* IEEE Transactions on Image Processing, vol. 15, no. 10, pages 2892–2903, Oct. 2006. (Cited on page 52.)

[Charbonnier 1997] P. Charbonnier, L. Blanc-Feraud, G. Aubert and M. Barlaud. *Deterministic edge-preserving regularization in computed imaging.* IEEE Transactions on Image Processing, vol. 6, no. 2, pages 298–311, Feb. 1997. (Cited on page 32.)

[Chen 1998] S.S. Chen, D.L. Donoho and M.A. Saunders. *Atomic decomposition by basis pursuit.* SIAM Journal on Scientific Computing, vol. 20, pages 33–61, 1998. (Cited on pages 129 and 130.)

[Chesneau 2010] C. Chesneau, J. Fadili and J.-L. Starck. *Stein block thresholding for wavelet-based image deconvolution.* Electronic Journal of Statistics, vol. 4, pages 415–435, 2010. (Cited on pages 27, 93 and 97.)

[Cohen 1992] A. Cohen, I. Daubechies and J.-C. Feauveau. *Biorthogonal bases of compactly supported wavelets.* Communications on Pure and Applied Mathematics, vol. 45, no. 5, pages 485–560, 1992. (Cited on pages 18, 28, 37, 56, 60, 106 and 137.)

[Coifman 1992] R. R. Coifman, Y. Meyer and V. Wickerhauser. *Wavelet Analysis and Signal Processing.* In Wavelet and their Applications, pages 153–178, 1992. (Cited on page 26.)

[Coifman 2001] R. Coifman, F. Geshwind and Y. Meyer. *Noiselets.* Applied and Computational Harmonic Analysis, vol. 10, no. 1, pages 27–44, 2001. (Cited on page 137.)

[Combettes 2005] P. L. Combettes and V. R. Wajs. *Signal Recovery by Proximal Forward-Backward Splitting.* Multiscale Modeling and Simulation, vol. 4, no. 4, pages 1168–1200, 2005. (Cited on page 135.)

[Daubechies 1992] I. Daubechies. Ten lectures on wavelets. Society for Industrial and Applied Mathematics, 1 édition, 1992. (Cited on page 27.)

[Dherete 2003] P. Dherete and B. Rouge. *Image de-blurring and application to SPOT5 THR satellite imaging.* In Proceedings of IEEE International Conference on Geoscience and Remote Sensing Symposium, volume 1, pages 318–320, Jul. 2003. (Cited on pages 32, 104, 107 and 123.)

[Do 2005] M.N. Do and M. Vetterli. *The contourlet transform: an efficient directional multiresolution image representation.* IEEE Transactions on Image Processing, vol. 14, no. 12, pages 2091–2106, Dec. 2005. (Cited on page 93.)

[Donoho 1994] D.L. Donoho and I.M. Johnstone. *Ideal Spatial Adaptation by Wavelet Shrinkage.* Biometrika, vol. 81, no. 3, pages 425–455, 1994. (Cited on page 25.)

[Donoho 1995a] D.L. Donoho. *De-noising by soft-thresholding.* IEEE Transactions on Information Theory, vol. 41, no. 3, pages 613–627, May 1995. (Cited on page 52.)

[Donoho 1995b] D.L. Donoho. *Nonlinear Solution of Linear Inverse Problems by Wavelet-Vaguelette Decomposition.* Applied and Computational Harmonic Analysis, vol. 2, no. 2, pages 101–126, 1995. (Cited on page 26.)

[Donoho 2006] D.L. Donoho. *Compressed sensing.* IEEE Transactions on Information Theory, vol. 52, no. 4, pages 1289–1306, Apr. 2006. (Cited on pages 128 and 129.)

[Durand 2003] S. Durand and J. Froment. *Reconstruction Of Wavelet Coefficients Using Total Variation Minimization.* SIAM Journal of Scientific Computing, vol. 24, pages 1754–1767, 2003. (Cited on pages 102, 103, 108, 109, 110 and 111.)

[Everett 1963] H. Everett. *Generalized Lagrange Multiplier Method for Solving Problems of Optimum Allocation of Resources.* Operations Research, vol. 11, no. 3, pages 399–417, 1963. (Cited on pages 24, 29 and 55.)

[Fletcher 2007] A.K. Fletcher, S. Rangan and V.K. Goyal. *On the Rate-Distortion Performance of Compressed Sensing.* In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, volume 3, pages 885–888, Apr. 2007. (Cited on page 137.)

[Gish 1968] H. Gish and J. Pierce. *Asymptotically efficient quantizing.* IEEE Transactions on Information Theory, vol. 14, no. 5, pages 676–683, 1968. (Cited on page 160.)

[Glowinski 1984] R. Glowinski. Numerical methods for nonlinear variation al problems. Springer-Verlag, 1984. (Cited on page 133.)

[Golub 1979] G.H. Golub, M. Heath and G. Wahba. *Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter.* Technometrics, vol. 21, no. 2, pages 215–223, 1979. (Cited on page 93.)

[Goyal 2008] V.K. Goyal, A.K. Fletcher and S. Rangan. *Compressive Sampling and Lossy Compression.* IEEE Signal Processing Magazine, vol. 25, no. 2, pages 48–56, Mar. 2008. (Cited on page 130.)

[He 2002] B. He, L.-Z. Liao, D. Han and H. Yang. *A new inexact alternating directions method for monotone variational inequalities.* Journal on Mathematical Programming, vol. 92, no. 1, pages 103–118, 2002. (Cited on page 136.)

[Jalobeanu 2003] A. Jalobeanu, L. Blanc-Féraud and J. Zerubia. *Satellite Image Deblurring Using Complex Wavelet Packets.* International Journal of Computer Vision, vol. 51, pages 205–217, 2003. (Cited on page 27.)

[Jayant 1972] N.S. Jayant and L.R. Rabiner. *The application of dither to the quantization of speech signals.* Bell System Technical Journal, vol. 51, pages 1293–1304, 1972. (Cited on page 104.)

[Jianwei 2009] M. Jianwei and F.X. Le Dimet. *Deblurring From Highly Incomplete Measurements for Remote Sensing.* IEEE Transactions on Geoscience and Remote Sensing, vol. 47, no. 3, pages 792–802, 2009. (Cited on page 130.)

[Kalifa 2003a] J. Kalifa and S. Mallat. *Thresholding Estimators for Linear Inverse Problems and Deconvolutions.* The Annals of Statistics, vol. 31, no. 1, pages 58–109, 2003. (Cited on page 26.)

[Kalifa 2003b] J. Kalifa, S. Mallat and B. Rouge. *Deconvolution by thresholding in mirror wavelet bases.* IEEE Transactions on Image Processing, vol. 12, no. 4, pages 446–457, Apr. 2003. (Cited on pages 24, 25, 26, 45, 46 and 93.)

[Kasner 1999] J.H. Kasner, M.W. Marcellin and B.R. Hunt. *Universal trellis coded quantization.* IEEE Transactions on Image Processing, vol. 8, no. 12, pages 1677–1687, Dec. 1999. (Cited on pages 52, 58 and 72.)

[Kawata 1972] T. Kawata. Fourier analysis in probability theory. Probability and mathematical statistics. Academic Press, 1972. (Cited on page 173.)

[Kingsbury 1998] N. Kingsbury. *The Dual-Tree Complex Wavelet Transform: A New Technique For Shift Invariance And Directional Filters.* In Proceedings of IEEE Digital Signal Processing Workshop, pages 319–322, 1998. (Cited on page 27.)

[Kuhn 1951] H.W. Kuhn and A.W. Tucker. *Nonlinear programming.* In Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, pages 481–492, Berkeley and Los Angeles, 1951. University of California Press. (Cited on pages 29, 58 and 73.)

[Labate 2005] W.Q. Labate D. amd Lim, G. Kutyniok and G. Weiss. *Sparse multidimensional representation using shearlets.* In Proceedings of SPIE Conference on Wavelet Applications in Signal and Image Processing, pages 254–262, 2005. (Cited on pages 27 and 93.)

[Lambert-Nebout 2000] C. Lambert-Nebout, C. Latry, G.A. Moury, C. Parisot, M. Antonini and M. Barlaud. *On-board optical image compression for future high-resolution remote sensing systems.* In Proceedings of SPIE Conference on Applications of Digital Image Processing, Jul. 2000. (Cited on page 107.)

[Li 1998] J. Li and R.M. Gray. *Text and picture segmentation by the distribution analysis of wavelet coefficients.* In Proceedings of International Conference on Image Processing, pages 790–794, 1998. (Cited on page 105.)

[Lier 2008] P. Lier, C. Valorge and X. Briottet. Imagerie spatiale : des principes d'acquisition au traitement des images optiques pour l'observation de la terre. Editions Cépaduès, 2008. (Cited on pages 9, 18, 33, 34, 46 and 93.)

[Lipshitz 1992] S. P. Lipshitz, R. A. Wannamaker and J. Vanderkooy. *Quantization and Dither: A Theoretical Survey.* Journal of the Audio Engineering Society, vol. 40, no. 5, pages 355–375, 1992. (Cited on pages 104, 106, 107, 108, 109, 110, 111, 112, 170 and 175.)

[Lustig 2007] M. Lustig, D.L Donoho and J.M. Pauly. *Sparse* MRI*: The application of compressed sensing for rapid* MR *imaging.* Magnetic Resonance in Medicine, vol. 58, no. 6, pages 1182–1195, Dec. 2007. (Cited on pages 129 and 138.)

[Mallat 1989] S.G. Mallat. *A theory for multiresolution signal decomposition: the wavelet representation.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 11, no. 7, pages 674–693, Jul. 1989. (Cited on pages 17 and 44.)

[Mallat 2008] S. Mallat. A wavelet tour of signal processing. Academic Press, 2008. (Cited on pages 18 and 93.)

[Neelamani 2004] R. Neelamani, Hyeokho C. and R. Baraniuk. *ForWaRD: Fourier-wavelet regularized deconvolution for ill-conditioned systems.* IEEE Transactions on Signal Processing, vol. 52, no. 2, pages 418–433, Feb. 2004. (Cited on pages 26, 93 and 97.)

[Ng 2010] M. K. Ng, P. Weiss and X. Yuan. *Solving Constrained Total-variation Image Restoration and Reconstruction Problems via Alternating Direction Methods.* SIAM Journal on Scientific Computing, vol. 32, no. 5, pages 2710–2736, Aug. 2010. (Cited on page 136.)

[Ortega 1998] A. Ortega and K. Ramchandran. *Rate-distortion methods for image and video compression.* IEEE Signal Processing Magazine, vol. 16, no. 6, pages 23–50, 1998. (Cited on page 90.)

[O'Sullivan 1986] F. O'Sullivan. *A Statistical Perspective on Ill-Posed Inverse Problems.* Statistical Science, vol. 1, no. 4, 1986. (Cited on page 25.)

[Parisot 2000a] C. Parisot, M. Antonini and M. Barlaud. *EBWIC: a low complexity and efficient rate constrained wavelet image coder.* In Proceedings of IEEE International Conference on Image Processing, volume 1, pages 653–656, 2000. (Cited on page 15.)

[Parisot 2000b] C. Parisot, M. Antonini, M. Barlaud, C. Lambert-Nebout, C. Latry and G. Moury. *On board strip-based wavelet image coding for future space remote sensing missions.* In Proceedings of IEEE International Conference on Geoscience and Remote Sensing Symposium, volume 6, pages 2651–2653, 2000. (Cited on pages 30 and 37.)

[Parisot 2001] C. Parisot, M. Antonini, M. Barlaud, S. Tramini, C. Latry and C. Lambert-Nebout. *Optimization of the joint coding/decoding structure.* In Proceedings of IEEE International Conference on Image Processing, volume 3, pages 470–473, 2001. (Cited on pages 28 and 60.)

[Parisot 2002] C. Parisot, M. Antonini and M. Barlaud. *Stripe-based MSE control in image coding.* In Proceedings of IEEE International Conference on Image Processing, volume 2, pages 649–652, 2002. (Cited on pages 27, 28, 29 and 30.)

[Patel 2009] V.M. Patel, G.R. Easley and D.M. Healy. *Shearlet-Based Deconvolution.* IEEE Transactions on Image Processing, vol. 18, no. 12, pages 2673–2685, Dec. 2009. (Cited on pages 93, 97 and 112.)

[Pesquet 2009] J.-C. Pesquet, A. Benazza-Benyahia and C. Chaux. *A SURE Approach for Digital Signal/Image Deconvolution Problems.* IEEE Transactions on Signal Processing, vol. 57, no. 12, pages 4616–4632, Dec. 2009. (Cited on page 27.)

[Ramani 2008] S. Ramani, T. Blu and M. Unser. *Monte-Carlo Sure: A Black-Box Optimization of Regularization Parameters for General Denoising Algorithms.* IEEE Transactions on Image Processing, vol. 17, no. 9, pages 1540–1554, Sep. 2008. (Cited on pages 93 and 145.)

[Roberts 1962] L. Roberts. *Picture coding using pseudo-random noise.* IRE Transactions on Information Theory, vol. 8, no. 2, pages 145–154, 1962. (Cited on page 104.)

[Rockafellar 1997] R.T. Rockafellar. Convex analysis. Princeton Mathematical Series, 1997. (Cited on pages 157 and 158.)

[Rudin 1992] L.I. Rudin, S. Osher and E. Fatemi. *Nonlinear total variation based noise removal algorithms.* Physica D: Nonlinear Phenomena, vol. 60, no. 1–4, pages 259–268, Nov 1992. (Cited on pages 32, 102 and 137.)

[Said 1996] A. Said and W.A. Pearlman. *A new, fast, and efficient image codec based on set partitioning in hierarchical trees.* IEEE Transactions on Circuits and Systems for Video Technology, vol. 6, no. 3, pages 243–250, Jun. 1996. (Cited on pages 19, 21, 22, 24, 52 and 132.)

[Schuchman 1964] L. Schuchman. *Dither Signals and Their Effect on Quantization Noise.* IEEE Transactions on Communication Technology, vol. 12, no. 4, pages 162–165, 1964. (Cited on pages 104 and 175.)

[Schulz 2009] A. Schulz, L. Velho and E.A.B. da Silva. *On the empirical rate-distortion performance of Compressive Sensing.* In Proceedings of IEEE International Conference on Image Processing, pages 3049–3052, Nov. 2009. (Cited on page 130.)

[Selesnick 2003] I.W. Selesnick and K.Y. Li. *Video denoising using 2D and 3D dual-tree complex wavelet transforms.* In Proceedings of Wavelet Applications in Signal and Image Processing, pages 607–618, 2003. (Cited on page 102.)

[Shannon 1948] C.E. Shannon. *A Mathematical Theory of Communication.* Bell System Technical Journal, vol. 27, no. 3, pages 379–423, 1948. (Cited on pages 57, 72, 90 and 138.)

[Shannon 1959] C.E. Shannon. *Coding theorems for a discrete source with a fidelity criterion.* In IRE International Convention Record, pages 142–163. Wiley Press, 1959. (Cited on pages 160 and 163.)

[Shapiro 1993] J. M. Shapiro. *Embedded image coding using zerotrees of wavelet coefficients.* IEEE Transactions on Signal Processing, vol. 41, no. 12, pages 3445–3462, Dec. 1993. (Cited on pages 19, 20, 24, 40, 41 and 52.)

[Sripad 1977] A.B. Sripad and D. Snyder. *A necessary and sufficient condition for quantization errors to be uniform and white.* IEEE Transactions on Acoustics,

Speech and Signal Processing, vol. 25, no. 5, pages 442–448, 1977. (Cited on page 170.)

[Stamm 2011] M.C. Stamm and K.J.R. Liu. *Anti-forensics of digital image compression.* IEEE Transactions on Information Forensics and Security, vol. 6, no. 3, pages 1050–1065, 2011. (Cited on pages 105, 106, 107, 108, 109, 110, 111 and 146.)

[Taubman 2000] D. Taubman. *High performance scalable image compression with EBCOT.* IEEE Transactions on Image Processing, vol. 9, no. 7, pages 1158–1170, Jul. 2000. (Cited on pages 19, 22, 23, 24, 52 and 132.)

[Thiebaut 2011] C. Thiebaut and R. Camarero. *CNES Studies for On-Board Compression of High-Resolution Satellite Images.* In B. Huang, editeur, Satellite Data Compression, pages 29–46. Springer, 2011. (Cited on page 17.)

[Tramini 1998] S. Tramini, M. Antonini, M. Barlaud and G. Aubert. *Quantization noise removal for optimal transform decoding.* In Proceedings of IEEE International Conference on Image Processing, volume 1, pages 381–385, Oct. 1998. (Cited on pages 15, 31, 32 and 102.)

[Tramini 1999] S. Tramini, M. Antonini, M. Barlaud and G. Aubert. *Optimal joint decoding/deblurring method for optical images.* In Proceedings of IEEE International Conference on Image Processing, volume 1, pages 381–385, 1999. (Cited on pages 30, 31 and 32.)

[Usevitch 1996] B. Usevitch. *Optimal bit allocation for biorthogonal wavelet coding.* In Proceedings of Data Compression Conference, pages 387–395, Mar. 1996. (Cited on pages 28, 32, 56, 72 and 78.)

[Uzawa 1958] H. Uzawa. Iterative methods for concave programming. Stanford University Press, 1958. (Cited on page 32.)

[Vanderkooy 1987] J. Vanderkooy and S.P. Lipshitz. *Dither in Digital Audio.* Journal of the Audio Engineering Society, vol. 35, no. 12, pages 966–975, 1987. (Cited on pages 54, 55, 171, 174 and 175.)

[Wallace 1992] G.K. Wallace. *The JPEG still picture compression standard.* IEEE Transactions on Consumer Electronics, vol. 38, no. 1, pages 18–34, Feb. 1992. (Cited on pages 16, 19, 127, 128 and 132.)

[Wannamaker 2000] R.A. Wannamaker, S.P. Lipshitz, J. Vanderkooy and J.N. Wright. *A theory of nonsubtractive dither.* IEEE Transactions on Signal Processing, vol. 48, no. 2, pages 499–516, Feb. 2000. (Cited on pages 54, 55, 71, 77, 171, 172, 173 and 174.)

[Weiss 2008] P. Weiss, L. Blanc-Féraud, T. Andre and M. Antonini. *Compression artifacts reduction using variational methods: Algorithms and experimental*

study. In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, pages 1173–1176, 2008. (Cited on pages 102, 103, 108, 109, 110 and 111.)

[Weiss 2009] P. Weiss, L. Blanc-Féraud and G. Aubert. *Efficient schemes for total variation minimization under constraints in image processing*. SIAM Journal on Scientific Computing, vol. 31, no. 3, pages 2047–2080, 2009. (Cited on page 136.)

[Widrow 1961] B. Widrow. *Statistical analysis of amplitude-quantized sampled-data systems*. Transactions of the American Institute of Electrical Engineers, vol. 79, no. 6, pages 555–568, 1961. (Cited on pages 169 and 170.)

[Wolf 1970] J. Wolf and J. Ziv. *Transmission of noisy information to a noisy receiver with minimum distortion*. IEEE Transactions on Information Theory, vol. 16, no. 4, pages 406–411, Jul. 1970. (Cited on pages 89 and 90.)

[Yeh 2005] P. Yeh, P. Armbruster, A. Kiely, B. Masschelein, G. Moury, C. Schaefer and C. Thiebaut. *The new CCSDS image compression recommendation*. In Proceedings of IEEE International Conference on Aerospace, pages 4138–4145, Mar. 2005. (Cited on page 38.)

Appendix A

Existence and uniqueness of optimal parameters

We detail here the existence and uniqueness of optimal parameters of the imaging chains addressed in Chapter 4.

A.1 Notions in optimization

We start by giving here some notions in optimization. The proofs of the following theorems can be found in [Rockafellar 1997].

Let $f : \mathbb{R}^N \to \mathbb{R}$ be a twice continously differentiable function and let $x = (x_1, x_2, \ldots, x_N)^T$ be a vector.

Theorem 1. *A point $x^* \in \mathbb{R}^N$ is a local minimum of f if there is an $\varepsilon > 0$ such that $f(x) \geq f(x^*)$ for all $x \in \mathbb{R}^N$ with $\|x - x^*\| < \varepsilon$.*

Corollary 1. *If $f(x) > f(x^*)$ for all $x \neq x^*$ with $\|x - x^*\| < \varepsilon$, then x^* is a strict local minimum of f.*

Theorem 2. *A point $x^* \in \mathbb{R}^N$ is a global minimum of f if $f(x) \geq f(x^*)$ for all $x \in \mathbb{R}^N$.*

Corollary 2. *If $f(x) > f(x^*)$ for all $x \neq x^*$, then x^* is a strict global minimum of f.*

Definition 1. *The gradient of f is the vector*

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}(x), \frac{\partial f}{\partial x_2}(x), \ldots, \frac{\partial f}{\partial x_N}(x) \right). \tag{A.1}$$

Definition 2. *The Hessian H of f is a $N \times N$ matrix defined as*

$$H_f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(x) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(x) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_N}(x) \\ \frac{\partial^2 f}{\partial x_1 \partial x_2}(x) & \frac{\partial^2 f}{\partial x_2^N}(x) & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_N}(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_N \partial x_1}(x) & \frac{\partial^2 f}{\partial x_2 \partial x_N}(x) & \cdots & \frac{\partial^2 f}{\partial x_N^2}(x) \end{bmatrix}. \tag{A.2}$$

Theorem 3. *If x^* is a local minimum, then the following conditions hold*

1. $\nabla f(x^*) = 0$,

2. $d^T H_f(x^*)d \geq 0$ *for all* $d \in \mathbb{R}^N$.

The Hessian matrix $H_f(x^*)$ is symmetric postive semi-definite, that is $x^T H_f(x^*)x \geq 0$ for any $x \in \mathbb{R}^N$ [Rockafellar 1997]. It is positive definite if we have a strict inequality: $x^T H_f(x^*)x > 0$ for any $x \in \mathbb{R}^N$.

Theorem 4. *For any* $x^* \in \mathbb{R}^N$, *if* $\nabla f(x^*) = 0$ *and* $H_f(x^*)$ *is positive definite, then* x^* *is a strict local minimum.*

We now introduce some results of convex optimization, which is a wide field of optimization.

Definition 3. *A set* $\Omega \subset \mathbb{R}^N$ *is said to be convex if, for all* x *and* y *in* Ω *and all* $t \in [0,1]$, *the following is verified*

$$tx + (1-t)y \in \Omega. \tag{A.3}$$

Definition 4. *A function* $f : \Omega \to \mathbb{R}$ *defined on a convex set* Ω *is said to be convex if for every* x *and* y *in* Ω *and all* $t \in [0,1]$, *we have*

$$f\left(tx + (1-t)y\right) \leq tf(x) + (1-t)f(y). \tag{A.4}$$

Theorem 5. *Let* f *be twice continously differentiable, then* f *is convex over a convex set* Ω *containing an interior point if and only if the Hessian matrix* H_f *is positive semi-definite in* Ω.

Theorem 6. *Let* f *be a convex function defined on a convex set* Ω. *Then, the set* X^* *where* f *achieves its minimum is convex. Furthermore, any local minimum is a global minimum.*

A.2 Optimal parameters of the on-ground chain

We now give a proof of existence and uniqueness of optimal parameters of the following problem

$$\inf_{\lambda_j>0,\,\Delta_j>0} \phi_\tau(\Delta_j, \lambda_j), \tag{A.5}$$

where

$$\begin{aligned}\phi_\tau(\{\Delta_j\},\{\lambda_j\}) = \sum_{j=0}^{J-1} \frac{\pi_j a_j \lambda_j^2}{(1+\lambda_j)^2}\sigma_{w_{x,j}}^2 + \frac{\pi_j a_j}{(1+\lambda_j)^2}\sigma_z^2 + \frac{\pi_j a_j \Delta_j^2}{12(1+\lambda_j)^2} \\ + \tau\left(\sum_{j=0}^{J-1} a_j R_j(\Delta_j) - R_c\right).\end{aligned} \tag{A.6}$$

To simplify the notations, we get rid of the constant R_c and the sum over j (as each subband is independent) in ϕ_τ, which now rewrites

$$\phi_\tau(\Delta_j, \lambda_j) = \frac{\pi_j a_j \lambda_j^2}{(1+\lambda_j)^2}\sigma_{w_{x,j}}^2 + \frac{\pi_j a_j}{(1+\lambda_j)^2}\sigma_z^2 + \frac{\pi_j a_j \Delta_j^2}{12(1+\lambda_j)^2} + \tau a_j R_j(\Delta_j). \tag{A.7}$$

Proposition 9. *Problem (A.5) admits an unique solution* $(\Delta_j^*, \lambda_j^*)$ *which verifies*

$$\lambda_j^* = \frac{\sigma_z^2}{\sigma_{w_{x,j}}^2} + \frac{\Delta_j^{*2}}{12\sigma_{w_{x,j}}^2} \tag{A.8}$$

$$\frac{\pi_j \Delta_j^*}{6(1+\lambda_j)^2} + \tau^* \frac{\partial R_j}{\partial \Delta_j}(\Delta_j^*) = 0 \tag{A.9}$$

Proof. To prove the existence and uniqueness of this solution, we propose to study the convexity of the function (A.7). We have

$$\frac{\partial \phi_\tau}{\partial \Delta_j}(\Delta_j, \lambda_j) = \frac{\pi_j a_j \Delta_j}{6(1+\lambda_j)^2} + \tau a_j \frac{\partial R_j}{\partial \Delta_j}(\Delta_j), \tag{A.10}$$

and

$$\frac{\partial^2 \phi_\tau}{\partial \Delta_j^2}(\Delta_j, \lambda_j) = \frac{\pi_j a_j}{6(1+\lambda_j)^2} + \tau a_j \frac{\partial^2 R_j}{\partial \Delta_j^2}(\Delta_j). \tag{A.11}$$

We also have

$$\begin{aligned}
\frac{\partial \phi_\tau}{\partial \lambda_j}(\Delta_j, \lambda_j) &= \pi_j a_j \sigma_{w_{x,j}}^2 \frac{\left(2\lambda_j(1+\lambda_j)^2 - 2(1+\lambda_j)\lambda_j^2\right)}{(1+\lambda_j)^4} - \pi_j a_j \sigma_z^2 \frac{2}{(1+\lambda_j)^3} \\
&\quad - \pi_j a_j \Delta_j^2 \frac{2}{12(1+\lambda_j)^3} \\
&= \frac{2\lambda_j \pi_j a_j \sigma_{w_{x,j}}^2}{(1+\lambda_j)^3} - \frac{2\pi_j a_j \sigma_z^2}{(1+\lambda_j)^3} - \frac{2\pi_j a_j \Delta_j^2}{12(1+\lambda_j)^3} \\
&= \frac{12\lambda_j \pi_j a_j \sigma_{w_{x,j}}^2 - 12\pi_j a_j \sigma_z^2 - \pi_j a_j \Delta_j^2}{6(1+\lambda_j)^3}
\end{aligned} \tag{A.12}$$

and

$$\begin{aligned}
\frac{\partial^2 \phi_\tau}{\partial \lambda_j^2}(\Delta_j, \lambda_j) &= \frac{12\pi_j a_j \sigma_{w_{x,j}}^2}{6(1+\lambda_j)^3} - \frac{12\lambda_j \pi_j a_j \sigma_{w_{x,j}}^2 - 12\pi_j a_j \sigma_z^2 - \pi_j a_j \Delta_j^2}{2(1+\lambda_j)^4} \\
&= \frac{4\pi_j a_j \sigma_{w_{x,j}}^2 (1+\lambda_j) - 12\lambda_j \pi_j a_j \sigma_{w_{x,j}}^2 + 12\pi_j a_j \sigma_z^2 + \pi_j a_j \Delta_j^2}{2(1+\lambda_j)^4} \\
&= \frac{4\pi_j a_j \sigma_{w_{x,j}}^2 - 8\lambda_j \pi_j a_j \sigma_{w_{x,j}}^2 + 12\pi_j a_j \sigma_z^2 + \pi_j a_j \Delta_j^2}{2(1+\lambda_j)^4}
\end{aligned} \tag{A.13}$$

Finally, we have

$$\frac{\partial^2 \phi_\tau}{\partial \lambda_j \partial \Delta_j}(\Delta_j, \lambda_j) = \frac{\partial^2 \phi_\tau}{\partial \Delta_j \partial \lambda_j}(\Delta_j, \lambda_j) = \frac{-2a_j \pi_j \Delta_j}{6(1+\lambda_j)^3} = \frac{-a_j \pi_j \Delta_j}{3(1+\lambda_j)^3}. \tag{A.14}$$

Using (A.10) and (A.12), we deduce the expressions (A.8) and (A.9) of the solution $(\Delta_j^*, \lambda_j^*)$ which satisfy the first-order conditions

$$\frac{\partial \phi_\tau}{\partial \Delta_j}(\Delta_j^*, \lambda_j^*) = 0, \tag{A.15}$$

$$\frac{\partial \phi_\tau}{\partial \lambda_j}(\Delta_j^*, \lambda_j^*) = 0. \tag{A.16}$$

To ensure that this solution exists and is unique, we study the convexity of ϕ_τ through its Hessian matrix H_{ϕ_τ}, which writes

$$H_{\phi_\tau}(\Delta_j, \lambda_j) = \begin{bmatrix} \frac{\partial^2 \phi_\tau}{\partial \Delta_j^2}(\Delta_j, \lambda_j) & \frac{\partial^2 \phi_\tau}{\partial \Delta_j \partial \lambda_j}(\Delta_j, \lambda_j) \\ \frac{\partial^2 \phi_\tau}{\partial \lambda_j \partial \Delta_j}(\Delta_j, \lambda_j) & \frac{\partial^2 \phi_\tau}{\partial \lambda_j^2}(\Delta_j, \lambda_j) \end{bmatrix}. \tag{A.17}$$

Since H_{ϕ_τ} is a 2×2 matrix, we can conclude that the function ϕ_τ is strictly convex if

$$\frac{\partial^2 \phi_\tau}{\partial \Delta_j^2}(\Delta_j, \lambda_j) > 0, \tag{A.18}$$

$$\frac{\partial^2 \phi_\tau}{\partial \lambda_j^2}(\Delta_j, \lambda_j) > 0, \tag{A.19}$$

and if the determinant of H_{ϕ_τ} is strictly positive

$$\det\left(H_{\phi_\tau}(\Delta_j, \lambda_j)\right) = \frac{\partial^2 \phi_\tau}{\partial \Delta_j^2}(\Delta_j, \lambda_j) \frac{\partial^2 \phi_\tau}{\partial \lambda_j^2}(\Delta_j, \lambda_j) - \left(\frac{\partial^2 \phi_\tau}{\partial \Delta_j \partial \lambda_j}(\Delta_j, \lambda_j)\right)^2 > 0. \tag{A.20}$$

The coding rate R_j is a monotonically decreasing positive function with respect to Δ_j [Shannon 1959], Δ_j being positive. Its limits are zero when Δ_j tends to infinity and infinity when Δ_j vanishes to zero [Gish 1968]. Its derivative $\frac{\partial R_j}{\partial \Delta_j}$ is negative and monotonically increasing, whose limits are minus infinity when Δ_j vanishes to zero and zero when Δ_j tends to infinity [Shannon 1959]. Still from [Shannon 1959], we have that $\frac{\partial^2 R_j}{\partial \Delta_j^2}$ is positive and monotonically decreasing. Since τ is positive, we deduce from (A.11) that

$$\frac{\partial^2 \phi_\tau}{\partial \Delta_j^2}(\Delta_j, \lambda_j) > 0, \quad \forall (\Delta_j, \lambda_j) \tag{A.21}$$

From equation (A.13), it is clear that $\frac{\partial^2 \phi_\tau}{\partial \lambda_j^2}$ is not always positive and we have

$$\frac{\partial^2 \phi_\tau}{\partial \lambda_j^2}(\Delta_j, \lambda_j) \begin{cases} > 0, & \text{if } 0 < \lambda_j < \lambda_j^h \\ = 0, & \text{if } \lambda_j = \lambda_j^h \\ < 0, & \text{otherwise,} \end{cases} \tag{A.22}$$

with

$$\lambda_j^h = \frac{1}{2} + \frac{12\sigma_z^2 + \Delta_j^2}{8\sigma_{w_{x,j}}^2}. \tag{A.23}$$

We need now to compute the determinant of the Hessian matrix H_{ϕ_τ}. Let us assume that $\frac{\partial^2 \phi_\tau}{\partial \lambda_j^2}(\Delta_j, \lambda_j)$ is strictly positive and let us define

$$g(\Delta_j, \lambda_j) = \frac{\pi_j a_j}{6(1+\lambda_j)^2} \frac{\partial^2 \phi_\tau}{\partial \lambda_j^2}(\Delta_j, \lambda_j) - \frac{a_j^2 \pi_j^2 \Delta_j^2}{9(1+\lambda_j)^6}. \tag{A.24}$$

Using equations (A.20) and (A.24), we have

$$\det\left(H_{\phi_\tau}(\Delta_j,\lambda_j)\right) = g(\Delta_j,\lambda_j) + \tau a_j \frac{\partial^2 R_j}{\partial \Delta_j^2}(\Delta_j)\frac{\partial^2 \phi_\tau}{\partial \lambda_j^2}(\Delta_j,\lambda_j) \tag{A.25}$$

Since $\tau a_j \frac{\partial^2 R_j}{\partial \Delta_j^2}(\Delta_j)$ is always strictly positive, we get the following inequality

$$\det\left(H_{\phi_\tau}(\Delta_j,\lambda_j)\right) > g(\Delta_j,\lambda_j), \tag{A.26}$$

such that if $g(\Delta_j,\lambda_j) > 0$ then we directly deduce that the Hessian matrix H_{ϕ_τ} is strictly positive and thus the function ϕ_τ is strictly convex. We have

$$\begin{aligned}
g(\Delta_j,\lambda_j) &= \left(\frac{\pi_j a_j}{6(1+\lambda_j)^2}\right)\left(\frac{\pi_j a_j\left(4\sigma_{w_{x,j}}^2 - 8\lambda_j\sigma_{w_{x,j}}^2 + 12\sigma_z^2 + \Delta_j^2\right)}{2(1+\lambda_j)^4}\right) \\
&\quad - \frac{a_j^2\pi_j^2\Delta_j^2}{9(1+\lambda_j)^6} \\
&= \frac{1}{3}\left(\frac{\pi_j^2 a_j^2\left(4\sigma_{w_{x,j}}^2 - 8\lambda_j\sigma_{w_{x,j}}^2 + 12\sigma_z^2 + \Delta_j^2\right)}{4(1+\lambda_j)^2} - \frac{a_j^2\pi_j^2\Delta_j^2}{3(1+\lambda_j)^6}\right) \\
&= \frac{1}{3}\left(\frac{3\pi_j^2 a_j^2\left(4\sigma_{w_{x,j}}^2 - 8\lambda_j\sigma_{w_{x,j}}^2 + 12\sigma_z^2 + \Delta_j^2\right) - 4a_j^2\pi_j^2\Delta_j^2}{12(1+\lambda_j)^6}\right) \\
&= \frac{1}{3}\left(\frac{12\pi_j^2 a_j^2\sigma_{w_{x,j}}^2 - 24\pi_j^2 a_j^2\lambda_j\sigma_{w_{x,j}}^2 + 36\pi_j^2 a_j^2\sigma_z^2 - a_j^2\pi_j^2\Delta_j^2}{12(1+\lambda_j)^6}\right).
\end{aligned} \tag{A.27}$$

From (A.27), we can conclude that $g(\Delta_j,\lambda_j) > 0$ if

$$12\pi_j^2 a_j^2\sigma_{w_{x,j}}^2 - 24\pi_j^2 a_j^2\lambda_j\sigma_{w_{x,j}}^2 + 36\pi_j^2 a_j^2\sigma_z^2 - a_j^2\pi_j^2\Delta_j^2 > 0, \tag{A.28}$$

that is, if

$$\lambda_j < \lambda_j^c, \tag{A.29}$$

where

$$\lambda_j^c = \frac{1}{2} + \frac{3}{2}\frac{\sigma_z^2}{\sigma_{w_{x,j}}^2} - \frac{\Delta_j^2}{24\sigma_{w_{x,j}}^2}. \tag{A.30}$$

Since $\lambda_j^c < \lambda_j^h$, we have from (A.22)

$$\frac{\partial^2 \phi_\tau}{\partial \lambda_j^2}(\Delta_j,\lambda_j) > 0, \quad \forall \Delta_j \,\text{and}\, \forall \lambda_j < \lambda_j^c, \tag{A.31}$$

which confirms the positivity hypothesis used to get inequality (A.26). We deduce that

$$\det\left(H_{\phi_\tau}(\Delta_j,\lambda_j)\right) > 0, \quad \forall (\Delta_j,\lambda_j) \in \mathbb{R}_+^* \times]0,\lambda_j^c[. \tag{A.32}$$

We can thus conclude that the function ϕ_τ is only convex locally on the convex domain $\mathbb{R}_+^* \times]0, \lambda_j^c[$.

From now, we set Δ_j to be equal to the optimal value Δ_j^*. Let us imagine that $\lambda_j^* > \lambda_j^c$, then we get that

$$\begin{aligned}
\frac{\sigma_z^2}{\sigma_{w_{x,j}}^2} + \frac{\Delta_j^{*2}}{12\sigma_{w_{x,j}}^2} &> \frac{1}{2} + \frac{3}{2}\frac{\sigma_z^2}{\sigma_{w_{x,j}}^2} - \frac{\Delta_j^{*2}}{24\sigma_{w_{x,j}}^2} \\
\frac{\sigma_z^2}{\sigma_{w_{x,j}}^2} + \frac{\Delta_j^{*2}}{12\sigma_{w_{x,j}}^2} &> \frac{\sigma_z^2}{\sigma_{w_{x,j}}^2} + \frac{1}{2}\frac{\left(\sigma_z^2 + \sigma_{w_{x,j}}^2\right)}{\sigma_{w_{x,j}}^2} - \frac{\Delta_j^{*2}}{24\sigma_{w_{x,j}}^2} \\
\frac{3\Delta_j^{*2}}{12\sigma_{w_{x,j}}^2} &> \frac{1}{2}\frac{\left(\sigma_z^2 + \sigma_{w_{x,j}}^2\right)}{\sigma_{w_{x,j}}^2} \\
\Delta_j^{*2} &> 2\left(\sigma_z^2 + \sigma_{w_{x,j}}^2\right), \qquad \text{(A.33)}
\end{aligned}$$

which is non-sense as it means that the optimal quantizing step would be greater than the standard deviation of the signal to quantize. In particular, note that $+\infty$ also verifies (A.33) although it completely cancels the signal. Condition (A.33) is also contradictory to the dithering hypothesis (4.12) that we made to develop our method, which comforts ourselves that this behavior never happens and that we always have $\lambda_j^* < \lambda_j^c$.

This result suggests that the point $(\Delta_j^*, \lambda_j^*)$ always lie in the strictly convex part of the function ϕ_τ.

By developing (A.23), we have

$$\begin{aligned}
\lambda_j^h &= \frac{1}{2} + \frac{12\sigma_z^2 + \Delta_j^2}{8\sigma_{w_{x,j}}^2} \\
&= \frac{1}{2} + \frac{3}{2}\frac{\sigma_z^2}{\sigma_{w_{x,j}}^2} + \frac{3}{2}\frac{\Delta_j^2}{12\sigma_{w_{x,j}}^2} \\
&= \frac{1}{2} + \frac{3}{2}\lambda_j^*, \qquad \text{(A.34)}
\end{aligned}$$

If the evaluate $\frac{\partial^2 \phi_\tau}{\partial \lambda_j^2}$ at the point $(\Delta_j^*, \lambda_j^*)$, we have from (A.22) and using the fact that $\lambda_j^* < \lambda_j^h$

$$\frac{\partial^2 \phi_\tau}{\partial \lambda_j^2}(\Delta_j^*, \lambda_j^*) > 0. \qquad \text{(A.35)}$$

Using (A.21), (A.32), (A.35) and Theorem 4, we deduce that the solution $(\Delta_j^*, \lambda_j^*)$ is a strict local minimum of the function ϕ_τ. If we look further at (A.10), we have

$$\frac{\partial \phi_\tau}{\partial \lambda_j}(\Delta_j^*, \lambda_j) \begin{cases} > 0, & \text{if } \lambda_j > \lambda_j^* \\ = 0, & \text{if } \lambda_j = \lambda_j^* \\ < 0, & \text{otherwise.} \end{cases} \qquad \text{(A.36)}$$

The derivative is strictly positive for any $\lambda_j > \lambda_j^*$, we deduce that

$$\phi_\tau(\Delta_j^*, \lambda_j) > \phi_\tau(\Delta_j^*, \lambda_j^*), \quad \forall \lambda_j > \lambda_j^*. \tag{A.37}$$

Since ϕ_τ is strictly convex on the domain $\mathbb{R}_+^* \times]0, \lambda_j^c[$ whose strict local minimum is λ_j^*, we have by Corollary 1

$$\phi_\tau(\Delta_j, \lambda_j) > \phi_\tau(\Delta_j^*, \lambda_j^*), \quad \forall \Delta_j \,\text{and}\, \forall \lambda_j \quad \text{with } 0 < \lambda_j < \lambda_j^c \quad \text{and } \lambda_j \neq \lambda_j^*. \tag{A.38}$$

Using (A.37) and (A.38), we have

$$\phi_\tau(\Delta_j, \lambda_j) > \phi_\tau(\Delta_j^*, \lambda_j^*) \quad \forall \Delta_j \,\text{and}\, \forall \lambda_j > 0 \quad \text{with } \lambda_j \neq \lambda_j^*, \tag{A.39}$$

which, by Corollary 2, concludes that the solution $(\Delta_j^*, \lambda_j^*)$ is the unique global minimum of the function ϕ_τ. □

We now have to deal with the numerical computation of the optimal parameters. Since the optimal regularizing parameter λ_j^* is expressed in closed-form, its computation is straightforward. The computation of the optimal quantizing step Δ_j^* is not direct as, for a given $\tau > 0$, we need to find a root of

$$g_\tau(\Delta) = \frac{\pi_j \Delta_j}{6\left(1 + \frac{\sigma_{w_{z,j}}^2}{\sigma_{w_{x,j}}^2} + \frac{{\Delta_j}^2}{12\sigma_{w_{x,j}}^2}\right)^2} + \tau \frac{\partial R_j}{\partial \Delta_j}(\Delta_j). \tag{A.40}$$

The monotony of the function g_τ is not easy to study since the term $\frac{\partial R_j}{\partial \Delta_j}$ is complex to evaluate. From our numerical experiments, we found out that the optimal quantizing step Δ_j^* always lies on a monotonically increasing part of the function g_τ. From this observation, we propose to use a binary search algorithm to compute this parameter. From (A.40), we see that Δ_j^* is function of τ. It seems reasonable to think that the higher τ is, the higher Δ_j^* needs to be for the function (A.40) to cross zero. This implies that the optimal quantizing step Δ_j^* can then be noted as a function of τ

$$\Delta_j^* = f(\tau), \tag{A.41}$$

where f is an increasing function. Consequently, from [Shannon 1959], we deduce that the coding rate R_j is a monotonically decreasing function with respect to τ. Using (4.38) and (A.41), we define

$$h(\tau) = \sum_{j=0}^{J-1} a_j R_j(f(\tau)) - R_c. \tag{A.42}$$

Then it seems clear that the function h is a monotonically decreasing function with respect to τ whose limits are infinity when τ vanishes to zero and $-R_c$ when τ tends to infinity. Its root τ^*, which verifies $h(\tau^*) = 0$, can then be computed using any root-finding algorithm. In our simulations, a binary search procedure will also be used.

A.3 Optimal parameters of the on-board chain

We now focus on the on-board chain and we give a proof of existence and uniqueness of optimal parameters in that case. The optimization problem still writes

$$\inf_{\lambda_j>0,\,\Delta_j>0} \phi_\tau(\Delta_j, \lambda_j), \tag{A.43}$$

where ϕ_τ is now given, after some simplifications, by

$$\phi_\tau(\Delta_j, \lambda_j) = \frac{\pi_j a_j \lambda_j^2}{(1+\lambda_j)^2}\sigma^2_{w_{x,j}} + \frac{\pi_j a_j}{(1+\lambda_j)^2}\sigma_z^2 + \frac{\pi_j a_j \Delta_j^2}{12} + \tau a_j R_j(\Delta_j). \tag{A.44}$$

Proposition 10. *Problem (A.43) admits an unique solution* $(\Delta_j^*, \lambda_j^*)$ *which verifies*

$$\lambda_j^* = \frac{\sigma_z^2}{\sigma^2_{w_{x,j}}} \tag{A.45}$$

$$\frac{\pi_j \Delta_j^*}{6} + \tau^* \frac{\partial R_j}{\partial \Delta_j}(\Delta_j^*) = 0 \tag{A.46}$$

Proof. To prove the existence and uniqueness of this solution, we also propose to study the convexity of the function (A.44). We have

$$\frac{\partial \phi_\tau}{\partial \Delta_j}(\Delta_j, \lambda_j) = \frac{1}{6}\left(\pi_j a_j \Delta_j\right) + \tau a_j \frac{\partial R_j}{\partial \Delta_j}(\Delta_j), \tag{A.47}$$

and

$$\frac{\partial^2 \phi_\tau}{\partial \Delta_j^2}(\Delta_j, \lambda_j) = \frac{1}{6}\left(\pi_j a_j\right) + \tau a_j \frac{\partial^2 R_j}{\partial \Delta_j^2}(\Delta_j). \tag{A.48}$$

We also have

$$\begin{aligned}
\frac{\partial \phi_\tau}{\partial \lambda_j}(\Delta_j, \lambda_j) &= \pi_j a_j \sigma^2_{w_{x,j}} \frac{\left(2\lambda_j(1+\lambda_j)^2 - 2(1+\lambda_j)\lambda_j^2\right)}{(1+\lambda_j)^4} - \pi_j a_j \sigma_z^2 \frac{2}{(1+\lambda_j)^3} \\
&= \frac{2\lambda_j \pi_j a_j \sigma^2_{w_{x,j}}}{(1+\lambda_j)^3} - \frac{2\pi_j a_j \sigma_z^2}{(1+\lambda_j)^3} \\
&= \frac{2\pi_j a_j \left(\lambda_j \sigma^2_{w_{x,j}} - \sigma_z^2\right)}{(1+\lambda_j)^3}
\end{aligned} \tag{A.49}$$

and

$$\begin{aligned}
\frac{\partial^2 \phi_\tau}{\partial \lambda_j^2}(\Delta_j, \lambda_j) &= 2\pi_j a_j \left(\frac{\sigma^2_{w_{x,j}}(1+\lambda_j)^3 - 3(1+\lambda_j)^2\left(\lambda_j \sigma^2_{w_{x,j}} - \sigma_z^2\right)}{(1+\lambda_j)^6} \right) \\
&= 2\pi_j a_j \left(\frac{\sigma^2_{w_{x,j}} - 2\lambda_j \sigma^2_{w_{x,j}} + 6\sigma_z^2}{(1+\lambda_j)^4} \right)
\end{aligned} \tag{A.50}$$

Finally, we have

$$\frac{\partial^2 \phi_\tau}{\partial \lambda_j \partial \Delta_j}(\Delta_j, \lambda_j) = 0. \tag{A.51}$$

In that case, the Hessian matrix is diagonal and therefore the convexity of the function ϕ_τ only depends on the sign of $\frac{\partial^2 \phi_\tau}{\partial \lambda_j^2}$. From (A.50), simply remark that

$$\frac{\partial^2 \phi_\tau}{\partial \lambda_j^2}(\Delta_j, \lambda_j) \begin{cases} > 0, & \text{if } 0 < \lambda_j < \lambda_j^c \\ = 0, & \text{if } \lambda_j = \lambda_j^c \\ < 0, & \text{otherwise,} \end{cases} \tag{A.52}$$

with

$$\lambda_j^c = \frac{1}{2} + 3\frac{\sigma_z^2}{\sigma_{w_{x,j}}^2}. \tag{A.53}$$

Since $\lambda_j^* < \lambda_j^c$, we can directly conclude that the point $(\Delta_j^*, \lambda_j^*)$ always lie in the strictly convex part of the function ϕ_τ. From (A.52), we deduce that

$$\frac{\partial^2 \phi_\tau}{\partial \lambda_j^2}(\Delta_j^*, \lambda_j^*) > 0. \tag{A.54}$$

Using (A.48), (A.54) and Theorem 4, we deduce that the solution $(\Delta_j^*, \lambda_j^*)$ is a strict local minimum of the function ϕ_τ. If we look further at (A.47), we have

$$\frac{\partial \phi_\tau}{\partial \lambda_j}(\Delta_j^*, \lambda_j) \begin{cases} > 0, & \text{if } \lambda_j > \lambda_j^* \\ = 0, & \text{if } \lambda_j = \lambda_j^* \\ < 0, & \text{otherwise.} \end{cases} \tag{A.55}$$

The derivative is strictly positive for any $\lambda_j > \lambda_j^*$, we deduce that

$$\phi_\tau(\Delta_j^*, \lambda_j) > \phi_\tau(\Delta_j^*, \lambda_j^*) \quad \forall \lambda_j > \lambda_j^*. \tag{A.56}$$

Since ϕ_τ is strictly convex on the domain $\mathbb{R}_+^* \times]0, \lambda_j^c[$ whose strict local minimum is λ_j^*, we have by Corollary 1

$$\phi_\tau(\Delta_j, \lambda_j) > \phi_\tau(\Delta_j^*, \lambda_j^*) \quad \forall \Delta_j \,\text{and}\, \forall \lambda_j \quad \text{with}\, 0 < \lambda_j < \lambda_j^c \quad \text{and}\, \lambda_j \neq \lambda_j^*. \tag{A.57}$$

Using (A.56) and (A.57), we have

$$\phi_\tau(\Delta_j, \lambda_j) > \phi_\tau(\Delta_j^*, \lambda_j^*) \quad \forall \Delta_j \,\text{and}\, \forall \lambda_j > 0 \quad \text{with}\, \lambda_j \neq \lambda_j^*, \tag{A.58}$$

which, by Corollary 2, concludes that the solution $(\Delta_j^*, \lambda_j^*)$ is the unique global minimum of the function ϕ_τ. □

In that case, the optimal parameters can be computed using the same numerical techniques than the ones we proposed for the on-ground chain.

A.4 Optimal parameters of the hybrid chain

We now focus on the hybrid chain. The optimization problem now writes

$$\inf_{\lambda_j>0,\,\mu_j,\,\Delta_j>0} \phi_\tau(\Delta_j,\mu_j,\lambda_j), \tag{A.59}$$

where ϕ_τ is now given, after some simplifications, by

$$\begin{aligned}\phi_\tau(\Delta_j,\lambda_j,\mu_j) = &\frac{\pi_j a_j\left(\lambda_j+\mu_j+\lambda_j\mu_j\right)^2}{(1+\lambda_j)^2(1+\mu_j)^2}\sigma^2_{w_{x,j}} + \frac{\pi_j a_j}{(1+\lambda_j)^2(1+\mu_j)^2}\sigma^2_{w_{z,j}}\\ &+\pi_j a_j\frac{\Delta_j^2}{12(1+\mu_j)^2} + \tau a_j R_j(\Delta_j).\end{aligned} \tag{A.60}$$

Proposition 11. *Problem (A.59) does not admit any solution.*

Proof. This result is slightly suprising but if we look at the first-order optimal conditions, we can remark than we are not able to find any acceptable solution. We have

$$\frac{\partial\phi_\tau}{\partial\lambda_j}(\Delta_j,\lambda_j,\mu_j) = \frac{2\pi_j a_j\sigma^2_{w_{x,j}}}{(1+\mu_j)^2}\frac{(\lambda_j\mu_j+\lambda_j+\mu_j)}{(1+\lambda_j)^3} - \frac{2\pi_j a_j\sigma_z^2}{(1+\mu_j)^2(1+\lambda_j)^3}, \tag{A.61}$$

and

$$\frac{\partial\phi_\tau}{\partial\mu_j}(\Delta_j,\lambda_j,\mu_j) = \frac{12\lambda_j\pi_j a_j\sigma^2_{w_{x,j}}\left(\lambda_j\mu_j+\lambda_j+\mu_j\right) - 12\pi_j a_j\sigma_z^2 - \pi_j a_j\Delta_j^2(1+\lambda_j)^2}{6(1+\mu_j)^3(1+\lambda_j)^2}. \tag{A.62}$$

We deduce that

$$\lambda_j^* = \frac{1}{1+\mu_j^*}\left(\frac{\sigma_z^2}{\sigma^2_{w_{x,j}}} - \mu_j^*\right), \tag{A.63}$$

and

$$\mu_j^* = \frac{1}{1+\lambda_j^*}\left(\frac{\sigma_z^2}{\sigma^2_{w_{x,j}}} + \frac{\Delta_j^2}{12\sigma^2_{w_{x,j}}}\left(1+\lambda_j^*\right)^2 - \lambda_j^*\right). \tag{A.64}$$

From Eq. (A.63), we have

$$\begin{aligned}1+\lambda_j^* &= \frac{\sigma_z^2}{\sigma^2_{w_{x,j}}(1+\mu_j^*)} + 1 - \frac{\mu_j^*}{1+\mu_j^*},\\ &= \left(\frac{\sigma_z^2}{\sigma^2_{w_{x,j}}}+1\right)\frac{1}{1+\mu_j^*},\end{aligned} \tag{A.65}$$

If we use Eq. (A.65) in (A.64), we obtain

$$\frac{\mu_j^*}{1+\mu_j^*}\left(\frac{\sigma_z^2}{\sigma_{w_{x,j}}^2}+1\right) = \frac{\sigma_z^2}{\sigma_{w_{x,j}}^2} - \frac{1}{1+\mu_j^*}\left(\frac{\sigma_z^2}{\sigma_{w_{x,j}}^2}-\mu_j^*\right) + \frac{\Delta_j^2}{12\sigma_{w_{x,j}}^2}\left(\frac{\sigma_z^2}{\sigma_{w_{x,j}}^2}+1\right)^2 \frac{1}{(1+\mu_j^*)^2}. \tag{A.66}$$

After some simplifications, we get that

$$\frac{\Delta_j^2}{12\sigma_{w_{x,j}}^2}\left(\frac{\sigma_z^2}{\sigma_{w_{x,j}}^2}+1\right)^2 \frac{1}{(1+\mu_j^*)^2} = 0, \tag{A.67}$$

which does not lead to a valid solution. Therefore, we cannot satisfy simultaneously

$$\frac{\partial \phi_\tau}{\partial \mu_j}(\Delta_j, \lambda_j, \mu_j) = 0, \tag{A.68}$$

and

$$\frac{\partial \phi_\tau}{\partial \lambda_j}(\Delta_j, \lambda_j, \mu_j) = 0. \tag{A.69}$$

□

We propose then to enforce the value of λ_j^* and we deduce the value of the other parameters by extension of Section A.2. Note that this choice is however suboptimal.

Appendix B

Review of non-subtractive and subtractive dithering techniques

B.1 Undithered system

We start this review by the presentation of an undithered system. This system is presented Fig. B.1.

Figure B.1: Undithered system.

The signal to quantize is noted x, y is the output of the system

$$y = Q(x), \tag{B.1}$$

where Q is an infinite mid-tread quantizer of step Δ. The transfer characteristics of this quantizer can be modeled as

$$Q(x) = \Delta \left\lfloor \frac{x}{\Delta} + \frac{1}{2} \right\rfloor, \tag{B.2}$$

where $\lfloor \rfloor$ is the floor function which returns the greatest integer less than or equal to its argument. Let ε be the global error (i.e. output minus input) of the system

$$\varepsilon = y - x = Q(x) - x = q(x), \tag{B.3}$$

where q is the quantization error function

$$q(x) = Q(x) - x. \tag{B.4}$$

If $-\frac{\Delta}{2} \leq x < \frac{\Delta}{2}$, then $y = 0$ and from (B.3) $\varepsilon = -x$. Similarly, if $\frac{\Delta}{2} \leq x < \frac{3\Delta}{2}$, then $y = \Delta$ and $\varepsilon = \Delta - x$. By extension, the conditional probability $p_{\varepsilon|x}$ can be expressed as [Widrow 1961]

$$\begin{aligned} p_{\varepsilon|x}(\varepsilon, x) &= \delta(\varepsilon - q(x)) \\ &= \delta\left(\varepsilon + x - \Delta \left\lfloor \frac{x}{\Delta} + \frac{1}{2} \right\rfloor\right) \\ &= \Pi_\Delta(\varepsilon) W_\Delta(\varepsilon + x), \end{aligned} \tag{B.5}$$

where Π_Δ is the rectangular window function

$$\Pi_\Delta(\varepsilon) = \begin{cases} \frac{1}{\Delta} & \text{if } -\frac{\Delta}{2} < \varepsilon \leq \frac{\Delta}{2}, \\ 0 & \text{otherwise} \end{cases} \tag{B.6}$$

and W_Δ is the sampling function

$$W_\Delta(\varepsilon) = \sum_{k=-\infty}^{+\infty} \delta(\varepsilon - k\Delta). \tag{B.7}$$

The probability density function p_ε of ε is then given by

$$\begin{aligned} p_\varepsilon(\varepsilon) &= \int_{-\infty}^{+\infty} p_{\varepsilon|x}(\varepsilon, x) p_x(x) dx \\ &= \Delta \Pi_\Delta(\varepsilon) \left[W_\Delta * p_x\right](-\varepsilon), \end{aligned} \tag{B.8}$$

where p_x is probability density function of x. From (B.8) it is clear that the rectangular function Π_Δ is wide enough such that at least one delta of the W_Δ function will contribute to the sum, and the position of this delta depends on p_x. Consequently, the global error of an undithered system cannot be made independent of the system input [Widrow 1961].

The characteristic function (defined as the Fourier transform of the probability density function) of ε writes

$$\begin{aligned} P_\varepsilon(u) &= \operatorname{sinc}(u) * [W_{\frac{1}{\Delta}}(-u) P_x(-u)] \\ &= \sum_{k=-\infty}^{+\infty} \operatorname{sinc}\left(u - \frac{k}{\Delta}\right) P_x\left(-\frac{k}{\Delta}\right) \\ &= \operatorname{sinc}(u) + \sum_{k=-\infty, k\neq 0}^{+\infty} \operatorname{sinc}\left(u - \frac{k}{\Delta}\right) P_x\left(-\frac{k}{\Delta}\right) \end{aligned} \tag{B.9}$$

where P_x is the characteristic function of x and

$$\operatorname{sinc}(u) = \begin{cases} \frac{\sin(\pi\Delta u)}{\pi\Delta u}, & \text{if } u \neq 0 \\ 1, & \text{otherwise} \end{cases}. \tag{B.10}$$

From equation (B.9), we see that the global error ε can be made uniformly distributed if the characteristic function P_ε is reduced to $\operatorname{sinc}(u)$. This gives rise to Theorem 7 [Lipshitz 1992].

Theorem 7. *The global error of an undithered system is not independent of the system input but can be made uniformly distributed if the characteristic function P_x of the system input verifies [Sripad 1977]*

$$P_x\left(\frac{k}{\Delta}\right) = 0, \quad \forall k \in \mathbb{Z}^*. \tag{B.11}$$

A direct consequence of Theorem 7 is that the global error of an undithered system is uniformly distributed if the probability density function of the system input can be expressed as the convolution product of uniform distributions. Note that the normal distribution also verifies this property if its standard deviation σ is large enough in front of the quantizing step [Vanderkooy 1987]

$$\sigma > \frac{\Delta}{2}. \tag{B.12}$$

To extend Theorem 7 to arbitrary probability density functions, a noise with a specific distribution can be inserted prior to quantizing. This noise can be either subtracted or not subtracted after the quantizing, giving two dithering systems: The non-subtractive and the subtractive dithering systems. Both systems are described in the next parts.

B.2 Non-subtractive dithering system (NSD)

We present here the extension of the undithered system to the case the system input is noised prior to quantizing [Wannamaker 2000]. This system is depicted Fig. B.2.

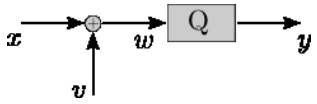


Figure B.2: Non-subtractive dithering system.

We keep the same notations than previously. The added noise, supposed to be independent of the source x, is noted v. The noisy signal w is now the input of the quantizer and we have

$$y = Q(w) = Q(x + v), \tag{B.13}$$

such that

$$\varepsilon = y - x = Q(x + v) - x = q(x + v) + v. \tag{B.14}$$

To study the statistical properties of the global error ε, the same technique than the one presented in Section B.1 can be used, except that the input of the quantizer is now w. Therefore, if $-\frac{\Delta}{2} \leq w < \frac{\Delta}{2}$, then $y = 0$ and $\varepsilon = -x$. Similarly, if $\frac{\Delta}{2} \leq w < \frac{3\Delta}{2}$, then $y = \Delta$ and $\varepsilon = \Delta - x$. By extension, the conditional probability $p_{\varepsilon|x}$ can be expressed as [Wannamaker 2000]

$$p_{\varepsilon|x}(\varepsilon, x) = \sum_{k=-\infty}^{+\infty} \delta(\varepsilon + x - k\Delta) \int_{-\frac{\Delta}{2}+k\Delta}^{\frac{\Delta}{2}+k\Delta} p_{w|x}(w, x)dw. \tag{B.15}$$

Using the fact that

$$p_{w|x}(w, x) = p_v(w - x), \tag{B.16}$$

where p_v is probability density function of the noise v, the conditional probability $p_{\varepsilon|x}$ rewrites

$$\begin{aligned} p_{\varepsilon|x}(\varepsilon,x) &= \sum_{k=-\infty}^{+\infty} \delta(\varepsilon+x-k\Delta) \int_{-\frac{\Delta}{2}+k\Delta}^{\frac{\Delta}{2}+k\Delta} p_v(w-x)dw \\ &= \sum_{k=-\infty}^{+\infty} \delta(\varepsilon+x-k\Delta) \int_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} p_v(w+k\Delta-x)dw \\ &= \sum_{k=-\infty}^{+\infty} \delta(\varepsilon+x-k\Delta) \int_{-\infty}^{\infty} \Delta\Pi_\Delta(w) p_v(\varepsilon+w)dw \\ &= W_\Delta(\varepsilon+x)[\Delta\Pi_\Delta * p_v](\varepsilon). \end{aligned} \tag{B.17}$$

We deduce the probability density function of the global error ε

$$\begin{aligned} p_\varepsilon(\varepsilon) &= \int_{-\infty}^{+\infty} p_{\varepsilon|x}(\varepsilon,x) p_x(x)dx \\ &= [\Delta\Pi_\Delta * p_v](\varepsilon)\,[W_\Delta * p_x](-\varepsilon). \end{aligned} \tag{B.18}$$

From (B.18), we see that for any choice of p_v (which is non-negative), the convolution product $\Pi_\Delta * p_v$ will give a function as wide as the rectangular window function. Similarly to the undither system, we deduce from this remark that the global error of a non-subtractive dithering system cannot be made independent of the system input [Wannamaker 2000].

The characteristic function of ε writes

$$\begin{aligned} P_\varepsilon(u) &= [\mathrm{sinc}(u)P_v(u)] * [W_{\frac{1}{\Delta}}(-u)P_x(-u)] \\ &= \sum_{k=-\infty}^{+\infty} \mathrm{sinc}\left(u-\frac{k}{\Delta}\right) P_v\left(u-\frac{k}{\Delta}\right) P_x\left(-\frac{k}{\Delta}\right). \end{aligned} \tag{B.19}$$

To be uniformly distributed, the characteristic function of ε must be reduced to $\mathrm{sinc}(u)$. If we admit that this is possible, we have for any $l \in \mathbb{Z}^*$

$$P_\varepsilon\left(\frac{l}{\Delta}\right) = \mathrm{sinc}\left(\frac{l}{\Delta}\right) = 0, \tag{B.20}$$

and from equation (B.19)

$$\begin{aligned} P_\varepsilon\left(\frac{l}{\Delta}\right) &= \sum_{k=-\infty}^{+\infty} \mathrm{sinc}\left(\frac{l}{\Delta}-\frac{k}{\Delta}\right) P_v\left(\frac{l}{\Delta}-\frac{k}{\Delta}\right) P_x\left(-\frac{k}{\Delta}\right) \\ &= P_x\left(-\frac{l}{\Delta}\right). \end{aligned} \tag{B.21}$$

By combining (B.20) and (B.21), we get that the global error of a non-subtractive dithering system can be made uniformly distributed if

$$P_x\left(\frac{l}{\Delta}\right) = 0, \quad \forall l \in \mathbb{Z}^* \tag{B.22}$$

which is not verified for arbitrary density probability functions. This gives Theorem 8 [Wannamaker 2000]

Theorem 8. *The global error of a non-subtractive dithering system is not independent of the system input and cannot be made uniformly distributed for arbitrary density probability functions p_x.*

From Theorem 8, we see that the independence of the global error cannot be obtained with this system. The moments of the global error can, however, be made independent of the system input for a certain class of dithering noise v [Wannamaker 2000]. The m-moment of the global error is given by

$$E[\varepsilon^m] = \int_{-\infty}^{+\infty} \varepsilon^m p_\varepsilon(\varepsilon) d\varepsilon, \tag{B.23}$$

and can also be expressed using its characteristic function P_ε [Kawata 1972]

$$E[\varepsilon^m] = \left(\frac{j}{2\pi}\right) P_\varepsilon^{(m)}(0), \tag{B.24}$$

where j is the imaginary number and $P_\varepsilon^{(m)}$ is the m derivative of the characteristic function P_ε. Let $G_v(u)$ be defined as

$$G_v(u) = \mathrm{sinc}(u) P_v(u), \tag{B.25}$$

such that, from (B.19), we have

$$E[\varepsilon^m] = \left(\frac{j}{2\pi}\right) \sum_{k=-\infty}^{+\infty} G_v^{(m)}\left(u - \frac{k}{\Delta}\right) P_x\left(-\frac{k}{\Delta}\right). \tag{B.26}$$

From equation (B.26), we deduce the following theorem [Wannamaker 2000]

Theorem 9. *The m-moment of the global error of a non-subtractive dithering system is independent of the system input if*

$$G_v^{(m)}\left(\frac{k}{\Delta}\right) = 0, \quad \forall k \in \mathbb{Z}^*. \tag{B.27}$$

If Theorem 9 is verified, the m-moment of the global error is given by

$$E[\varepsilon^m] = \left(\frac{j}{2\pi}\right) G_v^{(m)}(0), \tag{B.28}$$

which, by definition (B.25), is the same than the moment of the random variable composed by the sum of the dithering noise v plus a uniform random variable whose

probability density function is the rectangular window function (B.6). We, among others, then have [Wannamaker 2000]

$$E[\varepsilon] = E[v] \tag{B.29}$$

$$E\left[\varepsilon^2\right] = \sigma_v^2 + \frac{\Delta^2}{12}, \tag{B.30}$$

where σ_v is the standard deviation of the dithering noise v. Theorem 9 can be further developed

Theorem 10. *The m-moment of the global error of a non-subtractive dithering system is independent of the system input if*

$$P_v^{(l)}\left(\frac{k}{\Delta}\right) = 0, \quad \forall k \in \mathbb{Z}^* \text{ and } \forall l \in \{0, 1, 2, \ldots, m-1\}. \tag{B.31}$$

The proof of this theorem is adressed in [Wannamaker 2000]. If Theorem 10 is satisfied, an interesting corollary states that for a given m and for any n, the m-moment of the global error ε is independent from the n-moment of the system input x

$$E[\varepsilon^m x^n] = E[\varepsilon^m]E[x^n]. \tag{B.32}$$

A second interesting corollary is that Theorem 10 will be satisfied for any dithering noise v which is the sum of m uniformly distributed random variables [Wannamaker 2000]. In the thesis, we focus on dithering noise generated by a normal distribution. This type of dithering noise verifies Theorem 10 if its standard deviation σ_v is large enough in front of the quantizing step [Vanderkooy 1987]

$$\sigma_v > \frac{\Delta}{2}. \tag{B.33}$$

Although the moments independence may be sufficient for some applications, it is rarely exploited by image restoration algorithms which usually require stronger statistical properties such as signal independence. The latter can however be obtained using the subtractive dithering system described in the next part.

B.3 Subtractive dithering system (SD)

The subtractive dithering system is an extension of the non-subtractive scheme where the dithering noise v is substracted after quantizing. This system is depicted Fig. B.3.

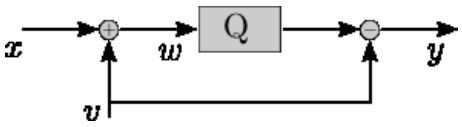


Figure B.3: Subtractive dithering system.

Using the same notations, we have

$$y = Q(w) - v = Q(x + v) - v, \tag{B.34}$$

such that

$$\varepsilon = y - x = Q(x + v) - (x + v) = q(x + v). \tag{B.35}$$

By analogy with equation (B.3), we see that the results of the subtractive dithering theory can be directly obtained from the undithered system theory by replacing x in part B.1 by $x + v$. We directly deduce [Lipshitz 1992]

$$p_\varepsilon(\varepsilon) = \Delta\Pi_\Delta(\varepsilon)\left[W_\Delta * p_x * p_v\right](-\varepsilon), \tag{B.36}$$

and

$$P_\varepsilon(u) = \mathrm{sinc}(u) + \sum_{k=-\infty, k\neq 0}^{+\infty} \mathrm{sinc}\left(u - \frac{k}{\Delta}\right) P_x\left(-\frac{k}{\Delta}\right) P_v\left(-\frac{k}{\Delta}\right). \tag{B.37}$$

In that case, the signal independence can be obtained if the following theorem is verified [Schuchman 1964]

Theorem 11. *The global error of a subtractive dithering system is independent from the system input and uniformly distributed between* $[-\frac{\Delta}{2}, \frac{\Delta}{2}]$ *if the characteristic function* P_v *of the dithering noise satisfies*

$$P_v\left(\frac{k}{\Delta}\right) = 0, \quad \forall k \in \mathbb{Z}^*. \tag{B.38}$$

which is true for any dithering noise generated by the sum of uniformly distributed random variables. Here again, the normal distribution verifies Theorem 11 if its standard deviation σ_v is large enough in front of the quantizing step [Vanderkooy 1987]

$$\sigma_v > \frac{\Delta}{2}. \tag{B.39}$$