Contributions en compression d'images médicales 3D et d'images naturelles 2D.

Yann Gaudeau

UFR Sciences et Techniques Mathématiques Informatique Automatique
Ecole Doctorale IAEM Lorraine
DFD Automatique et Production Automatisée

THÈSE

présentée pour l'obtention du

Doctorat de l'Université Henri Poincaré, Nancy 1

(Spécialité Automatique, Traitement du Signal et Génie Informatique)

par

Yann-Gaudeau

Contributions en compression d'images médicales 3D et d'images naturelles 2D.

Soutenue le 13 décembre 2006 devant la commission d'examen :

Rapporteurs :	Marc Antonini	Directeur de Recherche CNRS au laboratoire I3S
	Dominique Barba	Professeur émérite à l'Ecole Polytechnique de l'Université de Nantes
Examinateurs :	Cécile Germain-Renaud	Professeur à l'Université Paris-Sud
	Didier Wolf	Professeur à l'INPL
	David Brie	Professeur à l'Université Henri Poincaré, Nancy 1
	Jean Marie Moureaux	Maître de Conférences à l'Université Henri Poincaré, Nancy 1
	Michel Claudon	Professeur de médecine à l'Université Henri Poincaré, Nancy 1

Centre de Recherche en Automatique de Nancy
CRAN – CNRS UMR 7039

Table des matières

Table des figures

Liste des tableaux

Remerciements

Ce travail a été réalisé au Centre de Recherche en Automatique de Nancy. Je remercie son directeur ainsi que son prédécesseur Messieurs Alain Richard et Francis Lepage pour leur accueil.

Je tiens à exprimer toute ma reconnaissance à Monsieur Marc Antonini, directeur de recherche CNRS au I3S à Nice pour avoir accepté d'être rapporteur de ma thèse.

Je remercie également vivement Monsieur Dominique Barba, Professeur à l'Ecole Polytechnique de l'Université de Nantes, pour avoir bien voulu être rapporteur de ma thèse.

Je suis également très reconnaissant à Madame Cécile Germain-Renaud, Professeur à l'Université Paris-Sud, pour l'honneur qu'elle me fait de participer à mon jury de thèse.

Je tiens à remercier chaleureusement Monsieur Didier Wolf, Professeur à l'INPL pour avoir accepté de faire partie de mon jury de thèse.

Je remercie également vivement Monsieur Michel Claudon, Professeur de médecine au CHU de Nancy, pour l'honneur qu'il me fait de participer à mon jury de thèse.

Ce travail a été effectué sous la direction de Monsieur David Brie et Jean-Marie Moureaux, Professeur et Maître de conférences à l'université Henri Poincaré de Nancy. Je tiens à les remercier pour leur encadrement et les précieux conseils dont ils m'ont fait bénéficier.

Un grand merci à tous les doctorants et permanents du CRAN ainsi qu'au personnel du département Informatique de l'IUT Charlemagne pour l'ambiance qui règne dans ces deux lieux.

Je tiens à remercier encore une fois Jean-Marie pour ta disponibilité, ta simplicité, ta tolérance, ton honnêteté intellectuelle et ta gentillesse.

Merci à Ludovic Guillemot pour l'ensemble des travaux et échanges communs au cours de cette thèse. J'espère que tu trouveras le poste ou la fonction que tu mérites amplement.

Merci à Saïd Moussaoui pour ses étranges mixtures ! Je tiens tout particulièrement à saluer son engagement contre les restaurants qui mettent trois plombes pour servir !

Un grand merci à Phillipe Raffy pour la collaboration que nous avons réalisée sur le CAD et pour les publications qui en ont découlé.

Je remercie chaleureusement Monsieur Michel Lapeyre, aujourd'hui en poste au Centre Jean Perrin et initiateur de l'étude sur la compression avec pertes de scanners ORL.

Je remercie aussi vivement le Docteur Philippe Henrot et le docteur Pierre Graff pour leur disponibilité, et ce malgré leur emploi du temps très chargé.

Merci à toute l'équipe de l'ACI masse de données AGIR pour les échanges et le travail fourni dans le cadre de toutes les réunions aux quatre coins de l'Hexagone.

Merci à Monsieur Manuel Kalmecac, ex-général en chef des forces mexicaines de Nancy et vainqueur de la bataille du téléphone, pour ses éclats de rire discrets, son accent ensoleillé dans le bureau que l'on a partagé si longuement et tous les shows mexicains qu'il nous a proposé. A ce titre, merci à toute la communauté mexicaine de Nancy pour ses soirées et sa joie de vivre.

Merci à Monsieur V. Mazet pour des conseils avisés en tex, sa bisounourserie, et son art du badminton couplé à la danse classique. .

Merci Stéphane pour tes élans à me faire passer pour une personne raisonnable au CIES et pour avoir accepté la corruption qui t'a amené au sommet du conseil du CRAN..

Merci à J-P Planckert pour son art à titiller le "barbu".

Un grand merci à Philippe Jacques bâtiment S.A. pour ses conseils en matière de plomberie, en particulier lorsque l'on n'a pas les moyens de payer. Je salue également sa hot line dirigée de main de maître par Olivier Cervellin !

Merci et bonne chance à tous les autres thésards et ex thésards du labo : Fateh, Magalie, El-Hadi, Stéphane, Sinoë, Sébastien, Brian, Eric et Hicham. Sans oublier les ingénieurs CNAM, Christian et Jean-Marc !

Merci à tous les étudiants qui ont participé à cette thèse par l'intermédiaire de projet ou de stage : Audrey, Virginie, Nedia, Etienne, Hervé, Larry, Rodolphe, Bidon, Peste, Nicolas, Kamil, .

Merci à mon amour, à tous les autres potes, aux gens de Phi-Sciences et à la famille ... Cette partie qui est assez longue sera complétée en temps voulu.

Introduction générale

L'imagerie médicale est un domaine en plein essor, du fait du développement des technologies numériques. Elle permet une investigation de plus en plus fine des organes humains grâce à la mise à disposition de systèmes de radiologie de plus en plus performants. La contrepartie réside dans une quantité de données générée considérable qui peut rapidement saturer les systèmes conventionnels de transmission et de stockage. A titre d'exemple, le CHU de Nancy-Brabois produit des examens de type Pet-Scan où les données sont parfois produites à partir d'une acquisition allant de la tête aux pieds du patient. Ainsi, un seul de ces examens peut à lui seul dépasser aisément le Giga octet. Un autre exemple est la taille d'examens plus classiques comme l'IRM ou le scanner qui varie actuellement entre plusieurs dizaines et centaines de Méga octets. Pour un PACS[1] d'un service classique de radiologie, cette masse de données se chiffre à plusieurs Téra octets de données en une année. L'augmentation croissante et continue des capacités de stockage apporte une réponse partielle à ce problème mais demeure la plupart du temps insuffisante. La nécessité de compresser les images apparaît donc aujourd'hui incontournable. De plus, la compression présente un intérêt évident pour la transmission des images qui peut s'avérer délicate du fait des bandes passantes existantes limitées.

Actuellement, la compression dans un service de radiologie est toujours effectuée sans perte quand elle existe car elle constitue à ce jour le seul type de compression toléré par les médecins. En effet, la compression sans perte garantit l'intégrité des données et permet d'éviter les erreurs de diagnostic. Cependant, ce type de compression n'offre pas de réduction significative du volume de ces données. Dans ce contexte, la compression "avec pertes" maîtrisées peut être la réponse la plus appropriée, à condition bien entendu que les pertes n'affectent pas la qualité des images pour l'usage régulier des praticiens. Hier encore inenvisageable, l'idée d'une compression avec pertes semble aujourd'hui de mieux en mieux acceptée par les médecins, comme en témoigne par exemple, l'American College of Radiology (ACR) qui estime que les techniques de compression avec pertes peuvent être utilisées à des taux raisonnables, sous la direction d'un praticien qualifié, sans aucune réduction significative de la qualité de l'image pour le diagnostic clinique. L'un des principaux enjeux de ce manuscrit est donc de proposer une méthode de compression avec pertes efficace et acceptable visuellement pour les médecins.

Les méthodes actuelles de compression pour les images médicales 3D (pile de coupes 2D) sans et avec pertes les plus efficaces pour les images médicales exploitent la corrélation des images 3D à l'intérieur de la pile afin d'améliorer la performance de compression. Elles s'appuient sur une transformation décorrélant les trois dimensions : la **transformée en ondelettes 3D** (TO3D) et des algorithmes de quantification et de codage qui ont prouvé leur efficacité dans le cas des images 2D. Les **principaux codeurs actuels**, leurs fonctionnalités et leurs performances sont décrits dans le **chapitre 1**.

Dans cette logique, le **chapitre 2** présente une méthode[2] basée sur une TO3D dyadique avec une étape de quantification s'appuyant sur une **Quantification Vectorielle Algébrique avec Zone**

[1] Acronyme anglais de : Picture Archiving Communication System.

[2] Ces travaux ont été effectués en partie dans le cadre du projet AGIR (ACI Masse de Données) - http ://www.aci-agir.org/ -

Morte 3D (QVAZM 3D). La contribution de ce travail est la conception d'une zone morte multi-dimensionnelle pendant l'étape de quantification qui permet de prendre en compte les corrélations entre les voxels voisins. La zone morte vectorielle a pour but de seuiller les vecteurs de faible norme afin de quantifier plus finement les vecteurs de coefficients significatifs. Les performances de notre méthode ont tout d'abord été évaluées sur une base de données comprenant quatre scanners et quatre IRM de différents organes. Cette base est couramment utilisée pour tester les algorithmes de compression d'images volumiques. Les résultats montrent une supériorité visuelle et numérique de notre méthode par rapport à plusieurs des meilleures méthodes. Par ailleurs, une étude sur la compression avec pertes avec deux spécialistes (un radiothérapeute et un radiologue du Centre Alexis Vautrin de Nancy) a été mise en place sur des scanners ORL. Elle confirme les bonnes performances visuelles de la QVAZM 3D.

En imagerie grand public, l'utilisation massive d'Internet a banalisé l'usage de la compression avec pertes. Les images médicales constituent un tout autre défi car les algorithmes 3D actuels de compression avec pertes sont encore très loin d'être utilisés dans la pratique. Pour envisager un jour une mise en production de la compression avec pertes dans un PACS, il est fondamental de faire définitivement accepter l'idée de la compression avec pertes maitrisées à la communauté médicale. Dans cette optique, le **chapitre 3** mesure **l'impact d'une compression avec pertes sur les performances d'un CAD** (outil d'aide à la décision) performant pour **la détection automatique de nodules pulmonaires et des embolies pulmonaires**. Ce travail original sur une population significative de 120 scanners de poumons a montré que la détection des nodules restait robuste à compression jusqu'à de très fort taux de compression.

La QVAZM 3D et son prédécesseur la QVAZM 2D sont des algorithmes de compression intra-bandes nécessitant une procédure d'allocation de débits qui peut s'avérer très coûteuse pour de grosses masses de données comme les images médicales volumiques. Le **dernier chapitre** de ce manuscrit est consacré à la **réduction de la complexité** de notre schéma de compression à travers des outils analytiques pour les images naturelles et médicales. La première allocation consacrée à la QVAZM 2D s'appuie sur un a**justement précis des courbes débit-distorsion** en les approximant par un simple **modèle exponentiel**. Les paramètres de ce modèle sont calculés à partir d'un nombre réduit d'appels aux données. Ces courbes exponentielles conduisent à une résolution analytique de l'allocation de débits négligeable en terme de complexité. Nous avons obtenu des performances proches (en termes de qualité) de celles obtenues par une approche lagrangienne mais avec une complexité calculatoire considérablement réduite. La seconde allocation proposée pour les images naturelles 2D et médicales 3D utilise des **modèles statistiques par blocs** pour modéliser les courbes débit-distorsion. En effet, les agglutinements de coefficients d'ondelettes significatifs se traduisent par une **distribution de la norme des vecteurs** dont la forme, comportant un mode très proche de zéro, **ne peut être modélisée sous l'hypothèse i.i.d**. Nous montrerons que la distribution de la norme des vecteurs de coefficients d'ondelettes se modélise avec une grande précision également par un **mélange de lois gamma**. Cette propriété a débouché sur la modélisation de la distribution conjointe des vecteurs de coefficients d'ondelettes par un mélange de gaussiennes généralisées (multidimensionnelles). Ce travail permet de déduire des modèles débit-distorsion débouchant sur une procédure d'allocation des débit conduisant à des résultats numériques quasi équivalents à une approche lagrangienne classique avec un coût calculatoire réduit puisqu'il n'y a presque pas d'appel aux données.

Chapitre 1

Etat de l'art en compression d'images médicales

1.1 Introduction

La tendance actuelle est à l'utilisation croissante d'images médicales digitalisées. La plupart des techniques modernes d'imagerie médicale produisent des données 3D (IRM, scanner, échographie, tomographie par émission de positons) et même 4D (IRM fonctionnelle, échocardiographie 3D dynamique). Certaines images sont intrinsèquement volumiques alors que d'autres correspondent au contraire à une succession d'images 2D (encore appelée pile d'images), à laquelle on ajoute une dimension supplémentaire, à savoir l'écart entre deux coupes successives. De fait, la majorité des images médicales produites de nos jours peuvent être vues comme des images à au moins trois dimensions.

La quantité importante de ces images volumiques que génère chaque jour un PACS dans un service classique de radiologie se chiffre à plusieurs Téra octets de données en une année. Par ailleurs, ces données doivent être conservées un certain nombre d'années pour répondre aux contraintes réglementaires actuelles. L'augmentation croissante et continue des capacités de stockage apporte une réponse partielle à ce problème mais demeure la plupart du temps insuffisante. La compression semble donc incontournable pour résoudre ce problème d'archivage [113]. De plus, elle présente un intérêt évident pour la transmission des images qui peut s'avérer délicate du fait des bandes passantes existantes limitées.

Actuellement, la compression dans un service de radiologie est toujours effectuée sans perte quand elle existe. Elle est réalisée par des standards comme JPEG sans perte dont la syntaxe est prévue dans le standard de format d'images médicales DICOM[1]. Ce type de compression avec une reconstruction exacte de l'image de départ, garantissant l'intégrité des données demeure le préféré des praticiens pour des raisons évidentes de diagnostic. Cependant il offre de faibles performances en terme de débit binaire. Les taux de compression (TC) potentiels varient environ de 2 à 8 suivant le contenu informatif de l'image et la méthode appliquée. L'origine de la préférence des médecins pour la compression sans perte par rapport à la compression avec pertes est, comme on l'a dit, d'éviter les erreurs médicales liées à une mauvaise reconstruction de l'image. En effet, le principal problème de la compression avec pertes pour les images médicales est dû au fait que des détails importants pourraient disparaître (d'autres pourraient éventuellement apparaître). Ces détails sont généralement des structures difficiles à discerner car elles entraînent de faibles changements de contraste. Ainsi par exemple, des images peuvent révéler des lésions à travers des détails potentiellement sensibles à la compression avec perte puisqu'ils sont petits et ont des bords faiblement définis (comme par exemple : des microcalcifications dans des mammographies, le contour d'un

[1] Digital Imaging and COmmunications in Medecine (http ://medical.nema.org/).

pneumothorax, un faible nodule d'une image pulmonaire, etc). Notons enfin que les erreurs liées à l'utilisation éventuelle de la compression avec pertes pourraient entraîner pour le médecin des problèmes réglementaires et juridiques importants. Il faudrait en effet prouver qu'une erreur de diagnostic ne vient pas d'un problème de la qualité de l'image reconstruite.

Pour autant, la compression avec pertes est plus que jamais à l'ordre du jour en imagerie médicale, et ce pour les raisons suivantes. Tout d'abord, les études bien que peu nombreuses ont montré [36], [90], [91] que les images médicales possédaient des tolérances à la compression avec perte. On définit cette tolérance comme le maximum de taux de compression pour lequel l'image compressée est jugée acceptable, tant pour l'interprétation humaine que pour celle assistée par ordinateur (CAD[2]). Ainsi par exemple, les radiographies de poitrine digitalisées sont très tolérantes à la compression [36] (au moins 40 : 1 pour une compression avec la méthode SPIHT 2D [98] par exemple). Les films d'os digitalisés sont moyennement tolérants (entre 20 : 1 et 40 : 1). Les images de scanner, IRM et d'échographie sont plus faiblement tolérantes (de 10 : 1 à 20 : 1). Les études sur les scanners ont indiqué que la précision du diagnostic était préservée jusqu'a un taux de compression de 10 : 1 à la fois pour le thorax [24] et le foie [44], et jusqu'à 20 : 1 pour le colon [119]. Enfin, citons une étude très récente sur l'impact de la compression avec pertes (par SPIHT et JPEG2000) des mammographies. Cinq radiologues expérimentés ont localisé et noté les agglutinements de microcalcifications et les masses dans 120 mammographies [82]. Il en ressort que la précision des données mesurées est préservée jusqu'à des taux de compression de 80 : 1. Notons que toutes ces études ont été effectuées sur des méthodes 2D. Malheureusement, aucune étude n'a encore déterminé un taux de compression référence unique pour un type d'images (une modalité et un organe) et une méthode de compression. Cependant, nous pouvons déjà constater les gains en taux de compression offerts par rapport à ceux de la compression sans perte. En pratique, les radiologues avec l'aide des techniciens ont la capacité d'ajuster le niveau de compression de leurs archives par rapport à la modalité et la nature de l'examen. L'American College of Radiology (ACR) recommande que les techniques de compression avec pertes soient utilisées, sous la direction d'un praticien qualifié, car il estime qu'il n'y a aucune réduction significative de la qualité de l'image pour le diagnostic clinique [1]. En second lieu, la compression avec perte en imagerie médicale présente également de nombreux avantages pour d'autres applications que le diagnostic. En effet, certaines applications comme par exemple la radiothérapie n'ont pas nécessairement besoin de la même précision que les applications de diagnostic. Nous pouvons donc raisonnablement envisager une compression avec pertes maîtrisées pour ce type d'application. Enfin, la flexibilité de certains codeurs actuels permettant de faire de la compression avec pertes vers du sans perte, donnent la possibilité aux praticiens d'avoir l'image sans perte si nécessaire à partir d'une image comprimée avec perte. On parle alors de transmission progressive des données. Cette technique permet par exemple de raffiner progressivement la résolution des données visualisées en transférant des données supplémentaires. Les images sont décodées dans une petite taille pour finir à leur taille d'origine. Ce mode de fonctionnement fait donc référence à une scalabilité de résolution. Le praticien peut ainsi naviguer rapidement dans une base (ou pile) d'images de résolution réduite et ne demander le raffinement que de certaines coupes. Une autre utilisation de la propriété de scalabilité, consiste à décompresser immédiatement les images à pleine résolution (mais qualité médiocre) puis à raffiner les pixels en ajoutant des détails. C'est le cas par exemple des régions d'intérêt (ROI) que le médecin pourra sélectionner dans l'image, régions qui pourront être évaluées avec une précision très fine (sans perte).

Cependant, si la compression avec pertes d'images médicales semble intéressante à bien des égards, il faut noter que l'évaluation de la qualité après compression est un problème complexe aggravé par la nature de l'enjeu lorsqu'il s'agit de diagnostic. De façon similaire aux images classiques, l'évaluation de l'image reconstruite après compression avec pertes se base sur des critères objectifs et subjectifs. Et même s'il n'existe pas d'approche unique de mesure de la qualité qui amène à un

[2] Computer-Aided Detection.

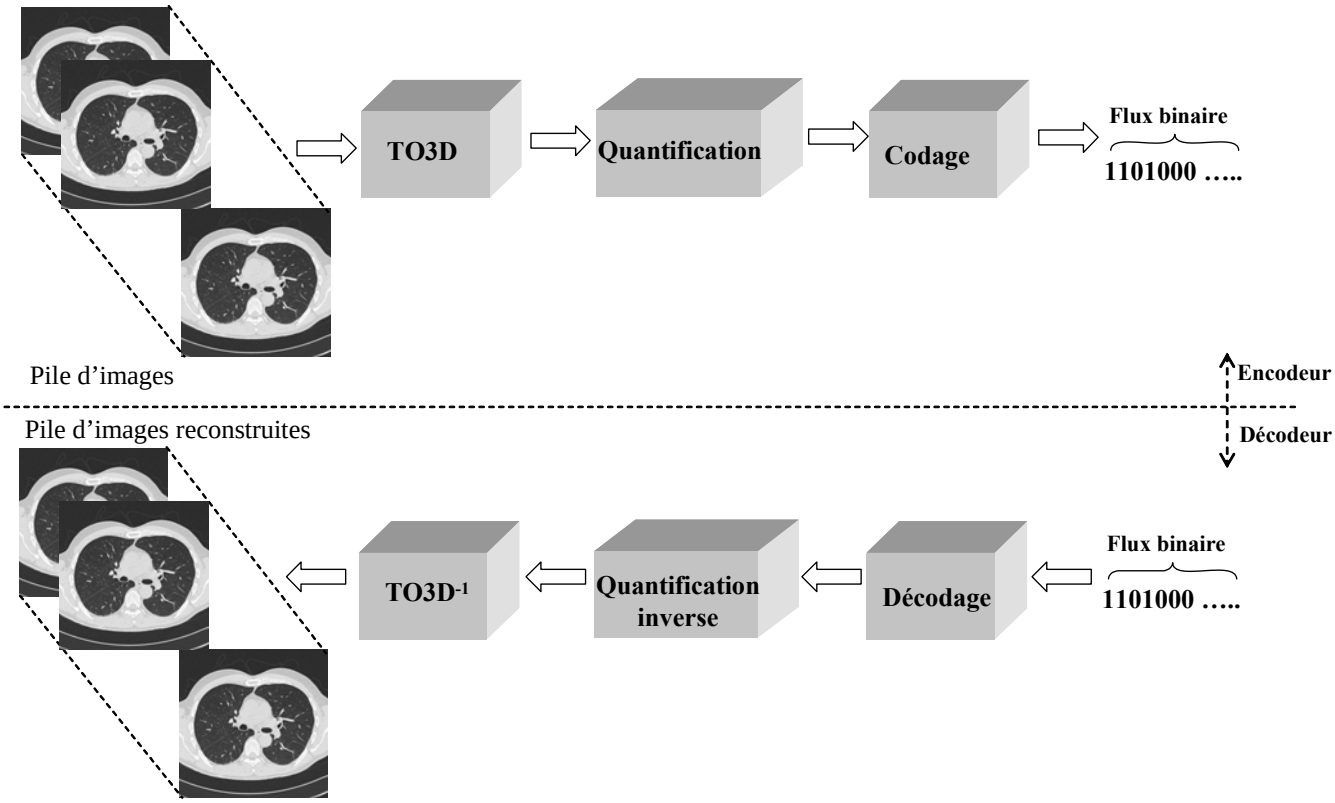

Fig. 1.1: Schéma de compression/décompression classique pour les images médicales volumiques 3D.

consensus sur le sujet [28], un certain nombre de critères statistiques ou subjectifs existent. Nous en parlerons plus longuement dans le paragraphe 3 et surtout dans le chapitre 3. D'ores et déjà, il est important de noter que l'appréciation de la qualité après reconstruction reste du ressort final des médecins.

Enfin, pour évaluer complètement une chaîne de compression, le troisième élément à prendre en compte avec le taux de compression et la qualité de l'image reconstruite est la complexité de l'application. Cet enjeu est encore plus important dans le cas des systèmes 3D. Si un temps long pour la compression des données peut être toléré, un décodage rapide est fondamental pour un accès rapide et efficace aux données. Il est exclu pour le médecin d'attendre trop longtemps pour voir l'image décompressée. Par ailleurs, dans le cas de la télémédecine, on doit faire face à une problématique de type temps réel. Autrement dit, on impose cette fois que la compression et la décompression soient rapides. Le facteur complexité est prédominant dans ce type d'application pour ce qui concerne la compression/décompression.

Ce chapitre a pour but de présenter les méthodes actuelles de compression 3D sans et avec pertes les plus efficaces pour les images médicales volumiques. L'ensemble de ce ces méthodes exploitent la corrélation des images 3D ou pile d'images 2D (très ressemblantes d'une image à l'autre) afin d'améliorer la performance de compression. Elles s'appuient sur une transformation décorrélant les trois dimensions : la transformée en ondelettes 3D (TO3D) et des étapes de quantification (pour la compression avec pertes)/codage qui ont prouvé leur efficacité dans le cas des images 2D. La figure 1.1 illustre le schéma de compression classique pour les images médicales volumiques.

Nous commencerons par introduire la problématique et les principaux outils d'évaluation de la qualité, les principes des trois étapes classiques de la compression d'images étant brièvement rappelés dans le paragraphe 3. Puis, nous présenterons de façon détaillée toutes les implantations

possibles de la transformée en ondelettes 3D au paragraphe 4. Le paragraphe 5 décrira les principaux algorithmes de codage actuels pour ce type d'images[3]. Le sixième paragraphe fera un bilan des performances de l'ensemble des méthodes évoquées au paragraphe 5.

1.2 Concepts et outils d'évaluation de la qualité

Comme nous l'avons déjà suggéré dans l'introduction de ce chapitre, l'évaluation de la qualité d'une image compressée avec pertes n'est pas une tâche facile, en particulier lorsqu'il s'agit d'images médicales. L'objet de ce paragraphe est de synthétiser les outils les plus utilisés dans ce domaine. Nous pouvons d'ores et déjà les répartir en trois catégories :

- **les mesures de distorsion objectives** telles que l'erreur quadratique moyenne (mean squared error - MSE - en anglais) ou le pic du rapport signal sur bruit (peak signal to noise ratio - PSNR - en anglais). Ces critères qui correspondent à l'analyse numérique des valeurs des pixels avant et après compression sont très généraux et ne vont pas toujours refléter la qualité de l'image reconstruite. Pour une image I de taille $L_x \times L_y$ pixels, on définit :

$$MSE = \frac{1}{L_x \times L_y} \left\| I - \hat{I} \right\|_2^2 \tag{1.1}$$

  où $\hat{I}$ représente l'image 2D après compression. Cette définition peut être étendue au voxel. De même, le rapport signal à bruit crête est défini de la manière suivante :

$$PSNR = 10 \log \frac{(2^b - 1)^2}{MSE} \quad dB \tag{1.2}$$

  où b représente le nombre de bits de codage.
  Ces mesures sont globales ; en aucun cas elles ne permettent de localiser une erreur dans l'image. Elles donnent une appréciation de la qualité d'ensemble de l'image et sont utilisées essentiellement pour comparer les méthodes ou les taux de compression. Elles sont simples d'utilisation et entrent également dans les processus d'optimisation des algorithmes de traitement (ici de compression) pour lesquels on cherche bien entendu à minimiser les dégradations totales (MSE) à taux de compression fixé. Par ailleurs, contrairement aux image naturelles [15] ou à la vidéo [16] , très peu d'études [85] ont été réalisées à notre connaissance pour proposer d'autres critères objectifs d'évaluation de la qualité spécifiques à l'imagerie médicale. Cela demeure un problème très ouvert qui ne sera pas abordé dans ce manuscrit.
- **les mesures de la qualité subjective** par des tests psychophysiques ou par des questionnaires avec notations réalisés par des professionnels de la radiologie :
  Dans ce cas, on demande à un ensemble de radiologues de noter la qualité des images comprimées (généralement sur une échelle de 1 à 5) selon que celles-ci leur permettent ou non d'effectuer un diagnostic (ou une mesure particulière d'organe). Les outils classiques d'analyse statistique (au minimum moyenne, écart-type) permettent ensuite d'interpréter les résultats et d'évaluer l'impact de la compression. Ces approches sont totalement ouvertes et ne reposent sur aucun standard (au contraire de l'approche suivante). Par ailleurs, contrairement aux image naturelles [15] ou à la vidéo [16], très peu d'études [85] ont été réalisées à notre connaissance pour proposer d'autres critères objectifs d'évaluation de la qualité spécifiques à l'imagerie médicale. Cela demeure un problème très ouvert qui ne sera pas abordé dans ce manuscrit

[3] Les algorithmes de codage sont présentés de manière un peu exhaustive dans ce chapitre, pouvant rendre la lecture un peu fastidieuse. Le lecteur intéressé seulement par les résultats pourra omettre de lire ces descriptions.

– **les simulations et analyses statistiques** d'une application spécifique comme la précision du diagnostic dans certaines images :
A l'intérieur de cette dernière catégorie, les méthodologies basées sur les courbes ROC[4] ont dominé historiquement [28]. Elles reposent sur deux grandeurs essentielles : la sensibilité et la spécificité, et nécessitent la définition d'un standard doré permettant d'établir une vérité "vraie" de diagnostic dans une base d'images. Nous décrivons un peu plus loin la notion de standard doré. Quoiqu'il en soit, une fois le standard établi, on appelle Vrai Positif (VP) toute lésion correctement retrouvée par le radiologue dans la base à juger, et Faux Négatif (FN) toute lésion existante manquée par le radiologue dans la même base. De même, toute lésion détectée à tort (par rapport au standard doré) est un Faux Positif (FP). La sensibilité (ou encore Proportion de Vrais Positifs - PVP) est définie alors comme étant le pourcentage de vraies lésions détectées par rapport au nombre total de vraies lésions :

$$Sensibilité = PVP = \frac{\#VP}{\#VP + \#FN} \tag{1.3}$$

La notion de Vrai Négatif (VN) est difficile à mesurer excepté dans le cas d'une détection binaire : à savoir le cas où chaque image ne peut avoir qu'une seule lésion d'un seul type, ou aucune lésion. Dans ce cas, un VN est une image qui n'est pas un VP. La spécificité se définit alors comme :

$$Spécificité = \frac{\#VN}{\#VN + \#FP} \tag{1.4}$$

Par exemple, en vénérologie, un test sur un virus est dit très sensible si une très grande proportion des patients porteurs du virus a été détectée par le test. De manière complémentaire, ce même test est qualifié de très spécifique s'il se révèle négatif pour une grande majorité de patients sains. Les très bons tests sont évidemment ceux qui sont à la fois spécifiques et sensibles.
La proportion de Faux Positifs (PFP), c'est à dire de lésions détectées à tort, n'est autre que le complément de la spécificité, à savoir $1 - Spécificité$. Si l'on trace les couples (PVP,PFP) obtenus pour différents taux de compression, on obtient un nuage de points dont la courbe la plus représentative est par définition la courbe ROC. Celle-ci doit passer nécessairement par les points (0,0) et (1,1). Afin de tenir compte de l'incertitude de diagnostic des radiologues, on demande à ceux-ci de donner un indice de confiance (de 1 à 5) pour chacun de leurs diagnostics. On ne garde ainsi que les couples (PVP,PFP) dont les indices sont suffisants [28]. Néanmoins, il est important de noter que l'analyse de type ROC possède plusieurs inconvénients parmi lesquels le fait qu'elle ne prenne pas en compte la localisation des lésions. D'autre part, la notation établie par les médecins ne reflète pas une pratique courante de la médecine et peut donc être sujette à caution. Notons enfin, qu'il est possible de tracer des courbes dites FROC[5] [28] qui permettent d'étendre le cas de la détection binaire à celui d'une détection multiple (on autorise alors un nombre arbitraire de lésions par image, à charge pour le radiologue de les retrouver dans l'image traitée en fournissant pour chacune d'elle un indice de confiance). Ces courbes sont également utilisées pour évaluer les performances des systèmes de détection automatique (CAD).
Comme on l'a dit précédemment, la définition d'un standard doré permet de déterminer une vérité "vraie" de diagnostic dans une base d'images originales, permettant ainsi la comparaison avec le diagnostic effectué sur les mêmes images traitées (c'est à dire compressées). On distingue en particulier trois types de standard :

1. le standard doré de type consensus : Il s'effectue uniquement sur les images originales (donc non compressées). Il est atteint lors de la lecture des résultats issus de tous les radiologues ou

[4] Receiver operating characteristic.
[5] Free-Response receiver Operating Characteristic.

après discussion de ces résultats (certains radiologues peuvent alors modifier leur opinion et certaines images peuvent être retirées de la base). Il est simple à mettre en oeuvre et favorise les images originales au détriment des images comprimées.

2. le standard doré de type personnel : Dans ce cas, c'est la même personne qui juge l'image originale et l'image comprimée. L'appréciation sur l'original est considérée comme une référence parfaite. Il est davantage pénalisant pour la compression. En effet, si un artéfact apparaît par erreur sur l'image originale et disparaît du fait de la compression, son absence sera jugée à tort par la même personne comme une information manquante ! Un inconvénient supplémentaire est qu'il est fortement dépendant du radiologue. Cependant, il possède l'avantage de n'écarter aucune image de la base et s'avère efficace pour comparer l'impact de différents taux de compression ou les méthodes de compression entre elles.
3. le standard doré de type indépendant : Le standard est basé sur l'accord des membres d'un panel d'éminents radiologues indépendants. Il est très utilisé dans les études. Il possède cependant un des inconvénients du standard précédent, à savoir qu'il est très sensible aux erreurs aléatoires. D'autre part, si un radiologue s'entête à effectuer un diagnostic sur des pixels équivoques, alors que le reste des membres du panel a écarté ce phénomène, il va introduire un biais de mesure.

Hormis les courbes ROC ou FROC, une grande variété d'approches ont également été utilisées dans lesquelles les radiologues sont amenés à exécuter différentes tâches d'interprétation. Ils détectent et localisent une pathologie, réalisent des mesures de structures variées et font des recommandations pour le traitement délivré au patient. L'image médicale peut donc être évaluée dans sa capacité à contribuer à ces différentes fonctions. Pour des informations supplémentaires sur l'évaluation de la qualité, nous invitons le lecteur à consulter les très complètes références suivantes : [25] [28], [29] et [30]. A présent que nous disposons de la connaissance des principaux concepts et outils liés à l'évaluation de la compression, nous allons rappeler les différentes étapes de la chaîne de compression pour les images numériques.

1.3 Les trois étapes classiques en compression

Les méthodes de compression avec pertes d'images médicales actuelles suivent les 3 étapes classiques de compression d'images naturelles. Elles débutent pour la plupart par une réorganisation du contenu de l'image, afin de séparer les composantes importantes (au sens visuel), des composantes contenant peu d'information. Cette tâche est remplie par une transformation mathématique. Cette étape est suivie par la quantification qui dégrade de manière irréversible le signal. La dernière étape de codage (sans perte) produit le flux binaire.

1.3.1 Première étape : Transformation des données

Le but de la transformée dans un schéma de compression est double. En effet, en plus de réorganiser l'information, elle doit représenter les composantes importantes d'un signal avec le moins d'éléments possibles : c'est ce qu'on appelle donner une représentation creuse du signal ou, de manière équivalente, compacter l'énergie.

Une image est constituée de contours et de textures : ce sont ces structures qu'il faut privilégier. En particulier, ce sont les contours qui rendent la plupart du temps possibles l'interprétation d'une image naturelle. Si l'on considère une image à niveaux de gris comme une fonction de deux variables, contours et textures correspondent à des variations brutales, voire à des discontinuités [27]. Les recherches menées en compression d'images sur la détermination de transformations performantes ont toutes pour but d'isoler et de caractériser de manière concise les contours et les textures.

Le but d'une transformation est de projeter un signal sur une base de fonctions dont les propriétés sont adaptées à la nature et aux caractéristiques du signal que l'on désire analyser. La projection est généralement orthogonale. Soit X un signal et T_X le signal projeté dans une base de fonction $(f_w)_{w\in\mathbb{R}}$, on a :

$$\forall w \in \mathbb{R},\ T_X(w) =< X, f_w >= \int_{\mathbb{R}} X(t) f_w(t) dt \tag{1.5}$$

avec $< ., . >$ le produit scalaire de l'espace L^2.

La première transformation mathématique employée pour analyser les signaux est la transformée de Fourier (TF). Celle-ci occupe une place centrale, notamment dans le domaine du traitement du signal, en raison de l'universalité liée au concept de fréquence et de son optimalité pour traiter les signaux stationnaires. En revanche, elle rend difficile l'analyse des signaux dits transitoires car les fonctions sur lesquelles s'effectuent la projection du signal sont définies sur $\mathbb{R}$. En effet, si l'on note TFx, la transformée de Fourier du signal X, on a :

$$TF_X(w) = \int X(t) e^{-j\pi wt} dt \tag{1.6}$$

La formule 1.6 montre que le calcul de $TF_X(w)$ nécessite de prendre en compte le signal X dans sa globalité : la transformée de Fourier ne permet pas de représenter des évenements concentrés dans le temps, ou dans l'espace, de manière équivalente pour les signaux 2D. Or la plupart des signaux réels tels que les sons ou les images ont des caractéristiques non stationnaires. Pour traiter ce type de signaux transitoires, on peut découper le signal en plusieurs morceaux sur lesquels sera appliquée une TF (ou une transformation dérivée) : on fait ici en quelque sorte l'hypothèse que le signal est "stationnaire par morceaux". C'est la méthode employée en particulier dans le standard de compression JPEG avec la transformée en cosinus discrète par bloc [107] [73]. Toutefois, même si cette stratégie donne de bons résultats, elle présente l'inconvénient de créer des artéfacts dûs au découpage artificiel en blocs à mesure que le taux de compression augmente

La seconde approche consiste à utiliser une base de fonctions qui ont une localisation temporelle (ou spatiale dans le cas des images) et fréquentielle. Obtenir une localisation fréquentielle consiste à utiliser une fonction g dont l'énergie est localisée sur un intervalle A**,** c'est-à-dire une fonction dont la valeur est négligeable en dehors d'un intervalle A. De cette manière, le calcul du produit scalaire (1.5) ne s'effectue que sur ce support compact :

$$< X, g >= \int X(t) g(t) dt \approx \int_A X(t) g(t) dt \tag{1.7}$$

La contribution du signal X dans le calcul (1.7) est localisée sur le support compact A. Une base de fonctions capable d'analyser en temps et fréquence tout le signal peut être obtenue par translation de la fonction g. La taille du support est un point très important car c'est lui qui détermine la résolution temporelle. Il existe de nombreuses recherches liées à la détermination de ce types de base et à la généralisation de ce principe [66]. Un des points culminants de ces travaux est la théorie des ondelettes qui sera présenté dans le paragraphe 1.4.

1.3.2 Deuxième étape : Quantification

Dans le schéma de compression, l'étape de quantification est celle qui dégrade de manière irréversible le signal. Elle est cependant d'une importance capitale dans la réduction du débit binaire. La quantification est une opération qui transforme un signal continu en un signal discret à l'aide d'un ensemble appelé dictionnaire [43][45]. Ce passage du continu au discret peut s'effectuer

échantillon par échantillon, dans ce cas on parlera de quantification scalaire (QS), ou bloc par bloc : c'est ce qu'on appelle la quantification vectorielle (QV). La présentation de la quantification vectorielle, et plus particulièrement de la quantification vectorielle algébrique est faite en détail dans le chapitre 2. On définit la fonction de quantification Q appliquée à un échantillon ou un bloc d'échantillons de la source X de la manière suivante :

$$\begin{array}{rcl} Q\colon \mathbb{R} & \rightarrow & C \\ X & \mapsto & Q\left(X\right) = \arg\min\limits_{Y \in C} \left\| X - Y \right\|_2^2 \end{array} \tag{1.8}$$

Le quantificateur Q associe à X le plus proche élément du dictionnaire C. L'objectif, lorsqu'on applique Q à une représentation creuse du signal (c'est-à-dire après transformation du signal), est de diminuer le nombre de bits nécessaires pour coder le signal. Les performances d'un quantificateur se mesurent en termes de débit binaire et de distorsion :

- Le nombre L d'éléments du dictionnaire détermine le débit binaire maximal $R_{\max}$:

$$R_{\max} = \lceil \log_2 \left(L\right) \rceil \text{ bits/échantillon} \tag{1.9}$$

- La distorsion correspond à l'erreur quadratique moyenne. Lorsque l'on cherche à déterminer le meilleur quantificateur au sens d'un critère quadratique, il est parfois intéressant de s'appuyer sur un modèle probabiliste permettant de modéliser la source X. Si l'on considère la source X comme un vecteur aléatoire de taille N et de distribution conjointe f_X, la distorsion D engendrée par la quantification s'écrit :

$$D = \int_{-\infty}^{+\infty} \left\| X - Q\left(X\right) \right\|_2^2 f_X\left(X\right) dX \tag{1.10}$$

La formule (1.10) est difficilement exploitable : il est souvent impossible de déterminer la distribution conjointe de l'ensemble du signal. Les modélisations classiques sont donc basées sur l'hypothèse d'une variable aléatoire i.i.d. de densité f, la formule de la distorsion devient alors :

$$D = \int_{-\infty}^{+\infty} \left(x - Q\left(x\right)\right)_2^2 f\left(x\right) dx \tag{1.11}$$

L'étape de quantification a pour but de déterminer pour un débit cible le dictionnaire qui minimisera la distorsion. On parle alors de compromis débit-distorsion.

Il existe plusieurs types de stratégies que nous allons brièvement présenter :

Quantificateurs scalaires uniformes (QSU) : Les représentants du dictionnaire sont répartis uniformément sur l'axe réel : la distance Δ entre deux représentants est constante. Le paramètre Δ s'appelle le pas de quantification, le quantificateur Q est alors déterminé par :

$$Q\left(x\right) = \left[\frac{x}{\Delta}\right] \tag{1.12}$$

avec [.] la fonction associant à un réel son plus proche entier et x un échantillon de X.

Ce type de quantification est très simple à mettre en oeuvre du fait du réglage d'un seul paramètre pour obtenir le débit souhaité. Par voie de conséquence, c'est le quantificateur le plus couramment utilisé.

Quantification scalaire avec zone morte : Les quantificateurs avec zone morte scalaire (zm) occupent une place centrale dans la compression par ondelettes car ils permettent d'éliminer

les coefficients non significatifs de faible magnitude au profit des coefficients significatifs qui sont alors quantifiés plus finement. On définit ce type de quantificateur de la manière suivante :

$$Q(X) = \begin{cases} 0 \text{ si } |x| < \delta \\ \left(\frac{x - signe(x) \times \left(\delta + \frac{\Delta}{2}\right)}{\Delta} \right) \text{ sinon} \end{cases} \tag{1.13}$$

avec δ le seuil de la zone morte.

L'intérêt de la zone morte réside dans le fait qu'elle permet de quantifier plus finement les coefficients significatifs de la source. Le point délicat de cette approche réside dans la détermination du seuil δ [87]. Le principe de la zone morte sera généralisé aux quantificateurs vectoriels dans le chapitre 2.

Nous ne pouvons pas clore ce paragraphe sans évoquer les travaux consacrés aux quantificateurs optimaux : le but dans ce cas est de déterminer les représentants du dictionnaire de manière à obtenir la distorsion totale la plus faible pour un débit donné. La méthode la plus connue est l'algorithme de Max-Loyd [43] [45]. Le principe consiste à déterminer les représentants du dictionnaire de sorte qu'ils aient la même probabilité d'apparition. La répartition obtenue est optimale car elle minimise la distorsion pour un débit fixe donné (voir formule (1.9)). Ces quantificateurs sont cependant plus complexes que les précédents car Δ varie. De plus ils semblent moins adaptés à la quantification des coefficients d'ondelettes car, de par leur principe, ils ont tendance à quantifier plus finement les coefficients non significatifs (les plus probables).

1.3.3 Troisième étape : Codage (sans perte)

Nous allons présenter dans ce paragraphe deux grandes familles de codeurs sans perte : les codeurs entropiques et les codeurs par plages. Ils sont utilisés dans une chaîne de compression sans perte, directement sur l'image de départ, ou après une transformée en ondelettes, comme nous le verrons au paragraphe 5. Ils sont également employés à la dernière étape de la chaîne de compression avec pertes (voire figure 1.1) afin d'exploiter les redondances présentes à la sortie du quantificateur.

Les codes entropiques sont basés sur la génération de mots dont la longueur dépend de la probabilité d'apparition des symboles de la source qu'il représente (on parle également de codes à longueur variable) : un grand nombre de bits sera utilisé pour coder un symbole peu probable tandis qu'un symbole redondant sera codé sur très peu de bits. Ce code doit en outre être uniquement décodable. La méthode pour générer de tels codes repose sur le principe du préfixe unique : aucun mot du code ne doit être le préfixe d'un autre mot. Par exemple, le code {0, 10, 110, 111} est un code préfixe alors que le code {0, 10, 101, 0101} ne l'est pas : 0 est le préfixe du mot 0101 et 10 est le préfixe de 101.

Il existe de nombreuses méthodes permettant de générer un code entropique, parmi lesquels le célèbre code de Huffman [50] [73]. Son principe est simple et repose sur la construction d'un arbre dont chaque branche est codée par 0 ou 1 selon la probabilité d'apparition des échantillons de la source. Cet algorithme admet cependant un inconvénient : il repose sur la connaissance complète de la source à coder (il existe cependant des versions adaptatives). D'autre part, il ne permet pas d'atteindre des débits inférieurs à 1 bit/échantillon.

Une autre approche du codage entropique ne présente pas le désavantage du codage d'Huffman : le codage arithmétique. La différence fondamentale entre les deux réside dans le fait qu'ici, les symboles ne sont pas codés séparément. En effet, c'est l'ensemble du message à transmettre qui est construit au fur et à mesure du traitement des différents éléments de la source [107]. Cette construction s'effectue en associant au message un nombre réel unique qui est transmis en binaire.

Comme nous l'avons déjà évoqué, les méthodes de codage entropique prennent en compte les probabilités d'apparition des symboles pour construire leurs codes. Des codes performants permettent d'approcher la borne inférieure que constitue l'entropie $\mathcal{H}$ (appelée également entropie

d'ordre zéro[6])de la source X [9] :

$$\mathcal{H}(X) = -\sum_{i=1}^{L} P(X = S_i) \log_2 P(X = S_i) \text{ bits/échantillon} \tag{1.14}$$

avec $\{S_1, ..., S_L\}$ l'ensemble des symboles de la source.

Toutefois, il est courant qu'en sortie du quantificateur, la source ne soit pas totalement décorrélée, c'est le phénomène de mémoire. Par exemple dans le cas des coefficients d'ondelettes, les coefficients de faible amplitude ont tendances à s'agglutiner. Les codeurs entropiques évoqués ci-dessus ne peuvent exploiter cette mémoire de la source quantifiée. Il existe néanmoins des codeurs capables d'exploiter les statistiques d'ordre supérieures : les codeurs entropiques adaptatifs et les codeurs par plages de zéros. Les codeurs par plages [52] (ou *run-length* en anglais) sont particulièrement adaptés pour coder les coefficients d'ondelettes. L'idée de base consiste à coder la longueur d'une série d'échantillons nuls plutôt que de coder chaque échantillon indépendamment. Le code run-length doit être associé ensuite à un codeur entropique classique.

Le paragraphe suivant se propose de d'expliquer le principe de la TO3D et ces différentes implémentations. Cette transformée est la première étape de la chaîne de compression proposée dans le chapitre 2.

1.4 La transformée en ondelettes : étape fondamentale du schéma compression

La transformée en ondelettes (TO) [32] est un outil largement utilisé en traitement du signal et d'image. L'une de ses applications principales est la compression d'images du fait sa capacité à compacter l'énergie sur un petit nombre de coefficients permettant un codage efficace de l'image. Elle correspond à la première étape de la chaîne de compression classique. Elle est couramment utilisé dans le cas des images 2D par application de filtres séparables dans les deux directions (axe x et y) [2]. La corrélation "temporelle" en imagerie médicale volumique a conduit à étendre cette transformée à la troisième dimension (axe z ou inter coupes) pour décorréler les trois dimensions. Cette transformée qui se nomme transformée en ondelettes 3D (TO3D) est maintenant admise comme référence en compression d'images volumiques [72], [103], [112], [118]. Le premier objectif de cette partie est de présenter les deux principales façons de calculer les coefficients d'ondelettes : par convolution classique et par lifting. Le second objectif de cette partie est de présenter les différentes implantations possibles de la TO3D et la difficulté à la rendre unitaire dans sa version entière.

1.4.1 La transformée en ondelettes par convolution classique

La transformée en ondelettes continue d'un signal à une dimension $X(t)$ fournit une décomposition du signal à différentes échelles grâce à l'utilisation d'une fonction mère $\psi_{u,s}$ dilatée d'un facteur s et translatée d'un facteur u [32]. Les fonctions $\Psi_{(u,s)}$ qui composent une base d'ondelettes $\left\{\Psi_{(u,s)}(t)\right\}_{(u,s)}$ sont définies de la manière suivante :

$$\Psi_{(u,s)} = \frac{1}{\sqrt{s}} \Psi\left(\frac{t-u}{s}\right) \tag{1.15}$$

[6]Notons qu'il existe des entropies d'ordre supérieur prenant en compte des blocs d'échantillons, plutôt que des échantillons isolés. De même, on définit des entropies dites conditionnelles qui tiennent compte des échantillons précédents l'échantillon (ou le bloc d'échantillons) courant. Ces différents types d'entropie permettent de mieux appréhender le phénomène de mémoire [43].

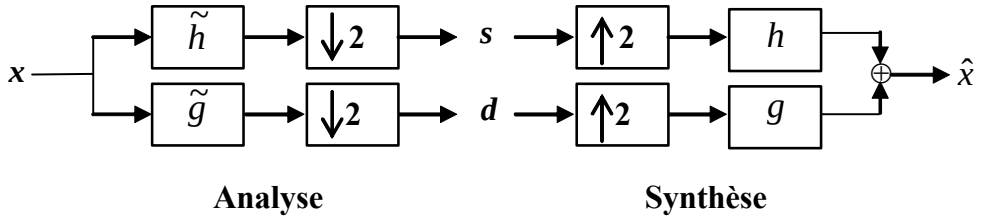

Fig. 1.2: Transformée en ondelettes 1D : analyse et synthèse.

On obtient de cette manière une analyse dont la résolution fréquentielle et temporelle est variable. C'est grâce à cette propriété que la théorie des ondelettes a connu un tel succès : faire varier u permet d'analyser localement le signal, et ce, à différentes échelles grâce à s. Cette échelle s permet de faire évoluer la résolution temps-fréquence de l'analyse en faisant varier la taille du support de $\Psi_{(u,s)}$ d'un rapport de $\frac{1}{s}$. Dans la suite, nous nous bornerons à définir et décrire les principales propriétés des ondelettes qui ont fait de cet outil un élément central dans les schémas de compression d'images actuels. Le lecteur intéressé par ce sujet est invité à consulter [6] [5] [66].

A partir de ces propriétés, l'idée de la TO est de décomposer une fonction quelconque $X(i)$ en une superposition d'ondelettes. Inversement, on peut reconstruire X à partir de sa fonction mère et des coefficients d'ondelettes obtenus lors de la décomposition.

Dans un cadre d'analyse multi-résolution, la transformée en ondelettes discrète peut être vue comme une décomposition du signal à différents niveaux de résolution (analyse multi-résolution) et peut être implantée en utilisant un couple de filtres discrets à travers un schéma classique d'analyse/synthèse (voir figure 1.2).

Dans l'étape d'analyse, le signal discret $x(n)$ est décomposé en 2 signaux, un signal d'approximation s en utilisant un filtre passe-bas $\widetilde{h}$ et un signal de détails d en se servant d'un filtre passe-haut $\widetilde{g}$. Cette décomposition est suivie par une décimation d'un facteur 2. Cela peut être représenté avec les équations suivantes en utilisant l'algorithme pyramidal proposé par Mallat [65] :

$$s(n) = \sum_k \widetilde{h}\,(2n-k)\,x(k) \tag{1.16}$$

$$d(n) = \sum_k \widetilde{g}\,(2n-k)\,x(k) \tag{1.17}$$

Le nombre total de coefficients de $s(n)$ et $d(n)$ est égal au nombre de coefficients dans $x(n)$. Dans l'étape de synthèse, $s(n)$ et $d(n)$ sont sur-échantillonnés et filtrés avec h et g. La somme des sorties du filtre produit le signal reconstruit $\widehat{x}(n)$. Si les filtres sélectionnés sont tels que

$$h\,(z)\,\widetilde{h}\left(z^{-1}\right) + g\,(z)\,\widetilde{g}\left(z^{-1}\right) = 2z^{-l} \tag{1.18}$$

$$h\,(z)\,\widetilde{h}\left(-z^{-1}\right) + g\,(z)\,\widetilde{g}\left(-z^{-1}\right) = 0 \tag{1.19}$$

alors le banc de filtres est parfaitement reconstruit avec un délai l . La figure 1.2 illustre cette implantation classique par banc de filtres de la transformée en ondelettes.

L'équation des coefficients de l'un des filtres de référence pour la compression d'image : le filtre biorthogonal 9.7 de Daubechies [2] (9 coefficients pour le filtre passe-bas $\widetilde{h}$ et 7 pour le filtre passe-haut $\widetilde{g}$). La référence [112] donne explicitement les coefficients d'autres filtres utilisables en imagerie médicale :

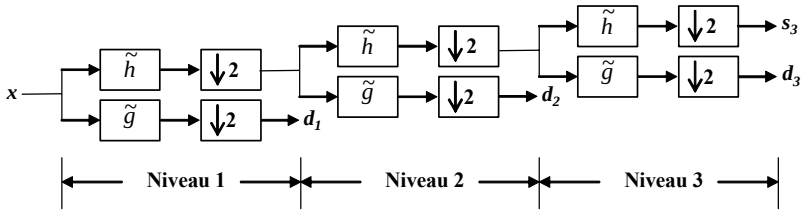

FIG. 1.3: Transformée en ondelettes dyadique sur 3 niveaux.

$$\widetilde{h} = \begin{pmatrix} 0,026749 \\ -0,016864 \\ -0,078223 \\ 0,266864 \\ 0,602949 \\ 0,266864 \\ -0,078223 \\ -0,016864 \\ 0,026749 \end{pmatrix} \qquad \widetilde{g} = \begin{pmatrix} 0,04536 \\ -0,028772 \\ -0,295636 \\ 0,557543 \\ -0,295636 \\ -0,028772 \\ 0,04536 \end{pmatrix} \tag{1.20}$$

Il est ensuite possible de décomposer $s(n)$ et $d(n)$. Dans un schéma dyadique, la sous-bande basse fréquence $s(n)$ est décomposée de manière récursive. La figure 1.3 montre une décomposition dyadique à 3 niveaux. Encore une fois, le nombre total de coefficients est égal au nombre d'échantillons d'entrée. Ainsi, les coefficients d'ondelettes peuvent être stockés dans un tableau de même taille que le signal d'entrée. A contrario, un schéma par paquets d'ondelettes autorise la récursivité de la décomposition sur la sous-bande haute fréquence $d(n)$.Dans ce paragraphe, nous venons de

présenter une méthode pour calculer les coefficients d'une transformée en ondelettes 1D à valeurs réelles ou entières. Les ondelettes 1D ont été successivement étendues aux images 2D, 3D et 4D en utilisant des filtres séparables dans chaque direction.

1.4.2 La transformée en ondelettes 3D dyadique

La TO a d'abord été étendue aux images 2D en appliquant de façon séparée les filtres d'analyse/synthèse sur les lignes et les colonnes [2]. Par conséquent, une transformée en ondelettes sur une image produit quatre sous images à la résolution inférieure :

- Une sous-bande basse-fréquence qui représente l'approximation de l'image à la résolution inférieure.
- Trois sous-bandes V, H et D correspondant au résultat du filtrage, respectivement, horizontal, vertical et diagonal (application des filtres verticaux et horizontaux). Ces sous-bandes contiennent les détails orthogonaux à la direction du filtrage. Par exemple un contour vertical correspondra à de forts coefficients dans la sous-bande V. Ce point apparaît de manière flagrante sur la figure 1.5.

Le succès des ondelettes en traitement d'images et particulièrement en compression est essentiellement basé sur deux points :

1. Les ondelettes permettent de représenter, à une échelle donnée, les variations brutales d'une image sur un faible nombre de coefficients. Cette propriété permet de délivrer une représentation creuse du signal comme le montre la figure 1.4.

2. Les images peuvent présenter des propriétés d'autosimilarités à différents niveaux d'échelles qui sont préservées par la transformée en ondelettes. C'est ce qu'on appelle le zoom multi-échelle [66]. Il existe donc des dépendances entre les représentations de l'image à différentes échelles. La figure 1.5 permet de se rendre compte de ces dépendances inter-échelles : les motifs présents dans l'image (b) se retrouvent bien sur les différentes échelles représentées sur la figure (b). Cette propriété est à la base des méthodes de compression de type arbres de zéro comme l'algorithme SPIHT [98] que nous décrirons un peu plus loin.

Soulignons enfin qu'il existe une troisième propriété liée à la localisation spatiale de la transformée en ondelettes et à l'utilisation de filtres orientés verticalement et horizontalement : le phénomène d'agglutinement. La figure 1.5 montre bien que les coefficients de forte amplitude ont tendance à se regrouper.

De façon similaire, la transformée en ondelettes 3D sur une image volumique (pile d'images 2D) peut être vue comme un produit séparable d'ondelettes 1D en les appliquant dans les trois directions spatiales [7], [112]. Après avoir été appliqués sur les lignes (x) et les colonnes (y), les filtres d'analyse/synthèse suivis d'une décimation 2 pour 1 sont appliqués le long de la troisième dimension (z). A la fin de la décomposition, 8 sous-bandes volumiques de résolution inférieure sont obtenus : l'image 3D basse-fréquence LLL et les 7 sous-bandes volumiques 3D de détails perdus. Les ondelettes 3D séparables fourniront une décorrélation égale des voxels[7] dans les trois directions. Notons que dans la suite du manuscrit, on omettra le qualificatif volumique quand on parlera des sous-bandes volumiques produites par la TO3D. On se limitera au terme sous-bande.

De façon identique au cas 1D, cette opération peut être appliquée autant de fois que le nombre de décompositions souhaité. Dans la transformée dyadique, chaque nouvelle décomposition est obtenue à partir d'une nouvelle transformée en ondelettes 3D sur la sous-bande basse-fréquence LLL. La figure 1.6 illustre la récursivité de la transformation en ondelettes 3D pour une transformée dyadique sur 3 niveaux.

Comme dans le cas 1D ou 2D, l'image volumique basse fréquence LLL_3[8] dans la figure 1.6 et les sous-volumes de détails perdus produits (coefficients d'ondelettes) par les différents niveaux de décomposition vont permettre de reconstruire l'image. Cette opération de synthèse consiste à remonter les résolutions en sens inverse de celui appliqué pendant l'analyse en appliquant à chaque fois les filtres de synthèse pour obtenir l'image 3D reconstruite.

Notons également qu'il est plus intéressant de réaliser une transformée en ondelettes 3D qu'une transformée en ondelettes 2D pour chaque image de la pile pour bien prendre en compte la corrélation entre les coupes. Il est donc acquis que l'utilisation d'une transformée en ondelettes 3D est toujours préférable à une succession de TO2D sauf dans le cas des images pour lesquelles la résolution sur le troisième axe est très faible. En effet, comme la résolution spatiale et par conséquent la dépendance entre les coupes est réduite, le bénéfice d'utiliser une transformation 3D décorrélant et implicitement un système de codage 3D décroît [103].

Nous pouvons également appliquer un nombre de niveaux de décomposition différent dans chaque direction spatiale. Cela permet d'adapter la taille de la pyramide d'ondelettes dans chaque direction spatiale dans le cas où la résolution spatiale est limitée. Par exemple, moins de niveaux sont requis le long de l'axe de coupe si le nombre de coupes ou la résolution le long de l'axe est limitée.

Notons aussi que des filtres différents peuvent être utilisés pour la transformation spatiale et temporelle. Dans [112], après avoir fixé l'utilisation du filtre 9/7 pour la transformation en x et y, les auteurs testent les performances de compression sur un scanner de genou en appliquant des filtres différents dans la direction z. Leur étude montre que le filtre 9/7 est bien adapté quand la

[7] Le voxel est un pixel en 3D (contraction de « volumetric pixel »).

[8] L et H signifie les filtres passe-haut et passe-bas respectivement. La direction du filtrage est dans l'ordre x, y et z. Le numéro à la fin correspond au niveau de décomposition.

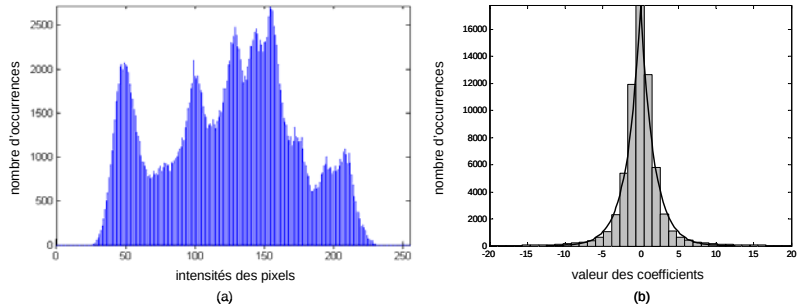

FIG. 1.4: (a) Histogramme des intensités des pixels d'une image - (b) Histogramme des coefficients d'ondelettes d'une sous-image (autre que la sous-image basse-fréquence).

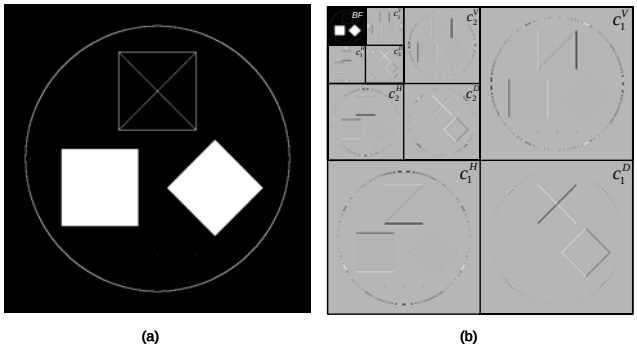

FIG. 1.5: Illustation des dépendances inter-échelles.

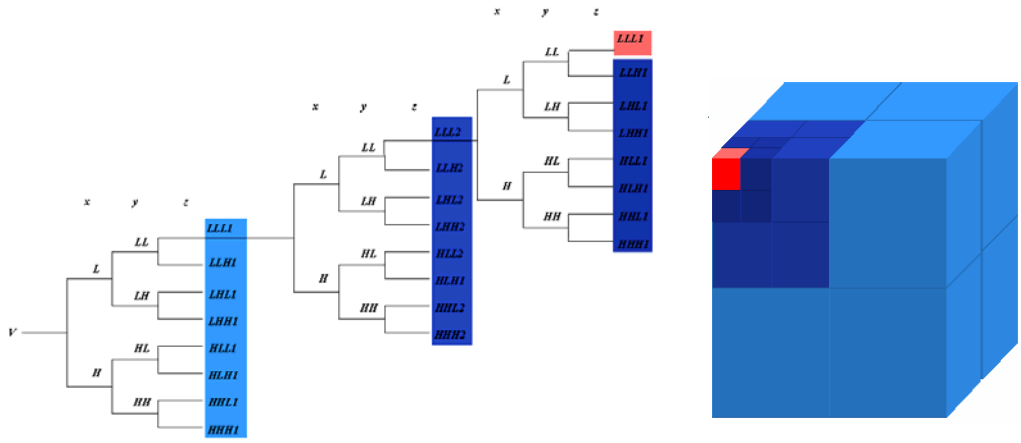

FIG. 1.6: Transformée en ondelettes 3D dyadique sur 3 niveaux : illustration de l'aspect volumique des sous-bandes.

distance inter-coupes est faible alors que le filtre de Haar est mieux indiqué pour une distance entre les coupes plus importante.

Nous avons présenté dans le paragraphe 1.4.1 la manière de calculer les coefficients d'ondelettes par une convolution classique. Cette méthode peut s'appliquer dans autant de dimensions que comporte l'image de départ mais elle produit toujours des coefficients réels. Or, un schéma classique de compression sans perte nécessite des coefficients entiers, la technique par lifting répondra à ce besoin.

1.4.3 Une implantation rapide et potentiellement entière de la TO : le lifting

Daubechies et Sweldens [33] ont utilisé un schéma appelé lifting pour calculer la TO discrète. Ils ont montré que n'importe quelle transformée en ondelettes pouvait être calculée à partir de ce schéma. Celui-ci permet de réduire le coût calculatoire par rapport à l'implantation par convolution car il utilise les redondances mathématiques entre le filtre passe-bas et le filtre passe-haut. Son second avantage est sa capacité à produire des coefficients d'ondelettes entiers. Cette fonctionnalité est fondamentale pour pouvoir envisager des méthodes de compression sans perte ou allant d'un schéma avec pertes vers du sans perte.

La transformée par lifting consiste dans un premier temps à séparer les échantillons pairs et impairs dans 2 tableaux différents. Si $x(n)$ est le signal d'entrée, ce découpage se note :

$$s^{(0)}(n) = x(2n) \tag{1.21}$$

$$d^{(0)}(n) = x(2n+1) \tag{1.22}$$

On applique ensuite une suite d'opérateurs de prédiction et de mise à jour pour obtenir les coefficients haute et basse fréquence de la transformée. Ces deux étapes forment un étage de lifting. Elles sont aussi appelées "pas de lifting dual" et "pas de lifting". Elles s'écrivent à l'itération :

$$d^{(i)}(n) = d^{(i-1)}(n) - \sum_k p^{(i)}(k) s^{(i-1)}(n-k) \tag{1.23}$$

$$s^{(i)}(n) = s^{(i-1)}(n) - \sum_k u^{(i)}(k) d^{(i)}(n-k) \tag{1.24}$$

où les coefficients $p^{(i)}(k)$ et $u^{(i)}(k)$ sont calculés en utilisant une factorisation Euclidienne de la matrice de polyphase [33]. Cette matrice est définie comme ceci :

$$P(z) = \begin{pmatrix} h_e(z) & g_e(z) \\ h_o(z) & g_o(z) \end{pmatrix} \tag{1.25}$$

où

$$\begin{array}{ll} h_e(z^2) = \frac{h(z)+h(-z)}{2} & g_e(z^2) = \frac{g(z)+g(-z)}{2} \\ h_o(z^2) = \frac{h(z)-h(-z)}{2z^{-1}} & g_o(z^2) = \frac{g(z)-g(-z)}{2z^{-1}} \end{array} \tag{1.26}$$

$h_e(z)$ et $h_o(z)$, respectivement $g_e(z)$ et $g_o(z)$, sont les composantes paires et impaires de la représentation polyphase des filtres de synthèse $h(z)$, respectivement $g(z)$. Si le déterminant de $P(z)$ est égal à 1, alors la paire de filtre (h, g) est complémentaire. Dans ce cas le théorème suivant s'applique [33] :

Théorème 1 : *Soit une paire de filtres (h, g) complémentaires, alors il existe toujours des polynômes de Laurent $s_i(z)$ et $r_i(z)$ pour $1 \leq i \leq m$ et une constante non nulle K tels que :*

$$P(z) = \prod_{i=1}^{m} \begin{pmatrix} 1 & r_i(z) \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ t_i(z) & 1 \end{pmatrix} \begin{pmatrix} K & 0 \\ 0 & \frac{1}{K} \end{pmatrix} \tag{1.27}$$

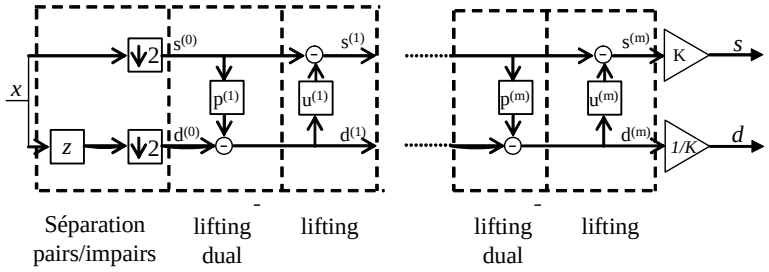

FIG. 1.7: Transformée en ondelettes utilisant un schéma de lifting (analyse).

Chaque matrice triangulaire 2×2 supérieure, respectivement inférieure correspond à une étape de prédiction (pas de lifting dual), respectivement de mise à jour (pas de lifting). Ainsi, les coefficients $p^{(i)}(k)$ et $u^{(i)}(k)$ des équations (1.23) et (1.24) sont déterminés par les polynômes de Laurent

$r_i(z)$ et $t_i(z)$. Le nombre d'étages de lifting m dépend à la fois de la longueur du filtre et de la décomposition elle-même. Notons également que la représentation de la matrice de polyphase $P(z)$ n'est pas unique. Ceci est dû à la non unicité de la solution de l'algorithme d'Euclide pour les polynômes de Laurent [33].

La figure 1.7 illustre un lifting à m étages. Cette figure est équivalente à la partie analyse de la figure 1.2. Les échantillons $s^{(m)}(n)$ deviennent les coefficients d'approximation $s(n)$ et les échantillons $d^{(m)}(n)$ deviennent les coefficients de détails. Le tout étant mis à l'échelle par un facteur de normalisation K :

$$s(n) = Ks^{(m)}(n) \tag{1.28}$$

$$d(n) = \frac{d^{(m)}(n)}{K} \tag{1.29}$$

La transformée inverse est équivalente à la partie synthèse de la figure 1.2. Les opérations sont simplement inversées :

$$s^{(m)}(n) = \frac{s(n)}{K} \tag{1.30}$$

$$d^{(m)}(n) = Kd(n) \tag{1.31}$$

Cette opération de mise à l'échelle est suivie par le lifting inverse et le lifting dual inverse :

$$s^{(i-1)}(n) = s^{(i)}(n) + \sum_k u^{(i)}(k)d^{(i)}(n-k) \tag{1.32}$$

$$d^{(i-1)}(n) = d^{(i)}(n) + \sum_k p^{(i)}(k)s^{(i-1)}(n-k) \tag{1.33}$$

Enfin, on reconstruit le signal x avec :

$$x(2n) = s^{(0)}(n) \tag{1.34}$$

$$x(2n+1) = d^{(0)}(n) \tag{1.35}$$

Dans la plupart des cas, la transformée en ondelettes produit des coefficients à valeurs réelles et bien que cela engendre en théorie une reconstruction parfaite de l'image originale, l'utilisation d'une précision arithmétique finie entraînera une reconstruction avec perte. Il est donc indispensable d'avoir une transformation produisant des entiers pour faire de la compression sans perte ou de la compression avec perte avec transmission progressive vers du sans perte. Traditionnellement ce type de transformation n'est pas facile à construire [14]. Cependant, la construction devient très simple avec le lifting [105]. En effet, il suffit d'ajouter une opération d'arrondi après chaque pas de lifting. Ainsi le lifting et le lifting dual deviennent :

$$d^{(i)}(n) = d^{(i-1)}(n) - \left\lfloor \left(\sum_k p^{(i)}(k) s^{(i-1)}(n-k) \right) + \frac{1}{2} \right\rfloor \tag{1.36}$$

$$s^{(i)}(n) = s^{(i-1)}(n) - \left\lfloor \left(\sum_k u^{(i)}(k) d^{(i)}(n-k) \right) + \frac{1}{2} \right\rfloor \tag{1.37}$$

où $.\lfloor . \rfloor$ désigne la partie entière inférieure.

Notons que bien que les équations (1.36) et (1.37) transforment des nombres entiers ($d^{(i-1)}(n)$, $s^{(i-1)}(n)$) en d'autres nombres entiers ($d^{(i)}(n)$, $s^{(i)}(n)$), les coefficients $p^{(i)}(k)$ et $u^{(i)}(k)$ ne sont pas nécessairement entiers. Ainsi calculer les coefficients transformés entiers peut requérir des opérations flottantes. Soulignons enfin que cette opération d'arrondi peut être modélisée par un bruit. Il a été montré qu'elle introduit une contribution additionnelle au bruit de quantification qui dégrade les performances débit/distorsion du système de codage. De plus, les arrondis sont responsables d'une oscillation du $PSNR$ le long de l'axe z, entraînant une qualité variable suivant l'endroit où l'on se situe dans les coupes. Nous invitons le lecteur intéressé à consulter [94] pour plus de détails. Le point important à retenir ici est que le bruit est proportionnel au nombre d'opérations d'arrondis, qui est lui-même fonction du nombre de décompositions et du nombre d'étages de lifting. Nous donnons quelques équations souvent utilisées pour le schéma de lifting entier en fixant pour l'instant le facteur de normalisation K à 1. Nous reviendrons plus en détails sur ce facteur de normalisation dans le paragraphe consacré à la transformation unitaire d'une TO3D. Les références [67], [118] donnent un large éventail de transformées par lifting performantes pour les images 3D. En voici trois :

Etape de lifting pour la transformation 5×3 (transformée sans perte de JPEG2000) [13], [107] :

$$\begin{cases} d(n) = x(2n+1) - \left\lfloor \frac{x(2n+2)+x(2n)}{2} + \frac{1}{2} \right\rfloor \\ s(n) = x(2n) + \left\lfloor \frac{d(n-1)+d(n)}{4} + \frac{1}{2} \right\rfloor \end{cases} \tag{1.38}$$

Etape de lifting pour le le filtre $S+P$ [99] :

$$\begin{cases} d^{(1)}(n) = x(2n+1) - x(2n) \\ s(n) = x(2n) + \left\lfloor \frac{d^{(1)}(n)}{2} \right\rfloor \\ d(n) = d^{(1)}(n) + \left\lfloor \frac{2}{8}\left(s(n-1) - s(n)\right) + \frac{3}{8}\left(s(n) - s(n+1)\right) + \frac{2}{8} d^{(1)}(n) + \frac{1}{2} \right\rfloor \end{cases} \tag{1.39}$$

Etape de lifting pour le le filtre $9/7$ F [33] :

$$\begin{cases} d^{(1)}(n) = d^{(0)}(n) - \left\lfloor \frac{203}{228}(s^{(0)}(n+1) + s^{(0)}(n)) + \frac{1}{2} \right\rfloor \\ s^{(1)}(n) = s^{(0)}(n) - \left\lfloor \frac{217}{4096}(d^{(1)}(n) + d^{(1)}(n-1)) + \frac{1}{2} \right\rfloor \\ d(n) = d^{(1)}(n) + \left\lfloor \frac{113}{228}(s^{(1)}(n+1) + s^{(1)}(n)) + \frac{1}{2} \right\rfloor \\ s(n) = s^{(1)}(n) + \left\lfloor \frac{1817}{4096}(d^{(1)}(n) + d^{(1)}(n-1)) + \frac{1}{2} \right\rfloor \end{cases} \tag{1.40}$$

1.4.4 Comment rendre la transformée en ondelettes 3D entière unitaire ?

Un problème classique lié à la transformée en ondelettes par lifting entière sans perte est la complexité pour réaliser une transformée unitaire. Les trois transformées (1.38), (1.39) et (1.40) entières données en exemple ne sont pas unitaires. Ce problème ne se pose pas pour certaines TO dyadiques réelles par convolution. Par exemple, une TO3D dyadique réelle séparable effectuée à partir du filtre biorthogonal 9/7 normalisé à 1 assurera une transformée presque unitaire : en d'autres terme l'erreur de quantification dans le domaine des ondelettes est égale à l'erreur quadratique moyenne dans le domaine spatial ou temporel. On peut noter cela de la manière suivante :

$$EQM = \frac{1}{L_x \times L_y \times L_z} \left\| I - \widehat{I} \right\|_2^2 \approx \frac{1}{L_x \times L_y \times L_z} \left\| W - \widehat{W} \right\| \tag{1.41}$$

où I est l'image 3D de départ de dimensions $L_x \times L_y \times L_z$, $\widehat{I}$ l'image 3D reconstruite, W l'image d'ondelettes 3D de même taille que I et $\widehat{W}$ sa version quantifiée.

Cette propriété est nécessaire pour permettre une bonne performance de compression avec perte. En calculant la norme L^2 des filtres passe-haut et passe-bas, le facteur de normalisation K peut être obtenu. Pour la TO2D, utiliser une TO par lifting entière n'est pas un problème puisque les facteurs de projections typiques pour avoir une transformation unitaire sont approximativement des puissances de 2 [99]. Ainsi, une approche simple pour rendre la TO2D entière approximativement unitaire est d'appliquer une normalisation en effectuant un simple décalage de bits. Le problème avec la TO3D dyadique utilisant un schéma de lifting entier est lié à la troisième dimension z. En d'autres termes, on se retrouve avec les facteurs de normalisation de la transformée en ondelettes dyadique entière 1D. Dans le cas 1D, les facteurs de normalisation pour les différentes sous-bandes ne sont pas tous des puissances de 2. Ainsi, un simple décalage de bits ne suffit pas pour calculer une transformation approximativement unitaire. Par exemple, une estimation acceptable pour de nombreux filtres consiste à normaliser avec $\sqrt{2}$ pour la sous-bande basse fréquence et $\frac{1}{\sqrt{2}}$ pour la sous-bande haute fréquence. La figure 1.8 donne le facteur de normalisation pour chaque sous-bande pour réaliser une transformée en ondelettes dyadique entière approximativement unitaire sur 4 niveaux dans le cas 1D. Nous pouvons remarquer que les facteurs de normalisation ne sont pas entiers empêchant une transformation unitaire entier vers entier. Deux solutions ont été proposées pour rendre la TO3D unitaire : la normalisation par lifting et la normalisation par décalage de bits à travers une structure en paquets d'ondelettes.

1.4.4.1 Normalisation par lifting

Le dernier pas de lifting correspondant à la normalisation dans 1.27 représenté par la matrice $\begin{pmatrix} K & 0 \\ 0 & \frac{1}{K} \end{pmatrix}$ peut être décrit par un processus de normalisation avec 4 pas de lifting [33].

$$\begin{pmatrix} K & 0 \\ 0 & \frac{1}{K} \end{pmatrix} = \begin{pmatrix} 1 & K - K^2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -\frac{1}{K^2} & 1 \end{pmatrix} \begin{pmatrix} 1 & K - 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \tag{1.42}$$

Ainsi pour la transformée 9/7 F par lifting, on ajoute 4 pas de lifting à l'équation. Avec le facteur de normalisation $K = \frac{1177}{1024} = 1,149$; les équations 1.40 deviennent :

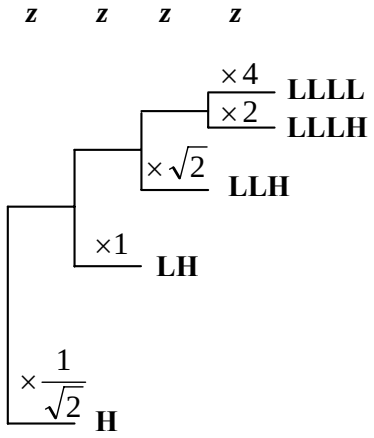

Fig. 1.8: Facteurs de normalisation pour réaliser une transformée en ondelettes entière 1D dyadique sur 4 niveaux approximativement unitaire.

$$\begin{cases} d^{(1)}(n) = d^{(0)}(n) - \lfloor \frac{203}{228}(s^{(0)}(n+1) + s^{(0)}(n)) + \frac{1}{2} \rfloor \\ s^{(1)}(n) = s^{(0)}(n) - \lfloor \frac{217}{4096}(d^{(1)}(n) + d^{(1)}(n-1)) + \frac{1}{2} \rfloor \\ d^{(2)}(n) = d^{(1)}(n) + \lfloor \frac{113}{228}(s^{(1)}(n+1) + s^{(1)}(n)) + \frac{1}{2} \rfloor \\ s^{(2)}(n) = s^{(1)}(n) + \lfloor \frac{1817}{4096}(d^{(1)}(n) + d^{(1)}(n-1)) + \frac{1}{2} \rfloor \\ d^{(3)}(n) = d^{(2)}(n) + s^{(2)}(n) \\ s^{(3)}(n) = s^{(2)}(n) + \lfloor \frac{153}{1024} d^{(3)}(n) + \frac{1}{2} \rfloor \\ d(n) = d^{(3)}(n) - \lfloor \frac{3563}{4096} s^{(3)}(n) + \frac{1}{2} \rfloor \\ s(n) = s^{(3)}(n) - \lfloor \frac{11}{64} d^{(3)}(n) + \frac{1}{2} \rfloor \end{cases} \tag{1.43}$$

Ces pas de lifting pour normaliser ne requièrent pas de mémoire supplémentaire qu'une normalisation avec des réels car seul un coefficient passe-bas et un coefficient passe-haut sont requis et ils sont disponibles en même temps du fait de l'imbrication des coefficients passe-bas et passe-haut dans le lifting.

1.4.4.2 Normalisation par décalage de bits par paquets d'ondelettes

Une autre structure de transformée pour s'assurer de l'unitarité est un arbre par paquets d'ondelettes. Ce type de structure est différent d'une transformée dyadique. La récursivité de la transformée ne s'applique pas uniquement sur la sous-bande basse fréquence. Elle peut s'appliquer sur toutes les sous-bandes de détails perdus. Ainsi, pour résoudre le problème, considérons tout d'abord la transformée 1D qui permet par un simple décalage de bits des coefficients d'ondelettes d'obtenir une transformée unitaire. La structure la plus simple répondant à nos besoins est une transformée par paquet d'ondelettes [22] sur 2 niveaux comme sur la figure 1.9. La figure 1.10 donne un exemple de structure 1D par paquet d'ondelettes à quatre niveaux. Les facteurs de normalisation sont tous des puissances de 2.

Le problème en 1D étant résolu, nous pouvons étendre cela au cas 3D [58] et [118] proposent d'appliquer la même transformée en ondelettes dyadique sur chaque coupe des données 3D volumiques. La figure 1.11 (a) illustre ces transformées 2D sur quatre coupes. Dans la dimension temporelle (axe z), nous appliquons une transformée par paquets d'ondelettes 1D comme décrit ci-dessus. Cela est montré dans le schéma 1.11 (b) où un arbre classique par paquets d'ondelettes à 2 niveaux est mis en oeuvre pour l'axe z. Les facteurs de normalisation pour chaque sous-bande sont aussi donnés. Ils sont la multiplication des facteurs de normalisation de la figure 1.9 et de la

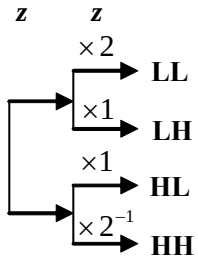

FIG. 1.9: Structure en paquet d'ondelettes 1D sur 2 niveaux permettant de rendre la transformation approximativement unitaire par un décalage implicite des coefficients d'ondelettes entiers.

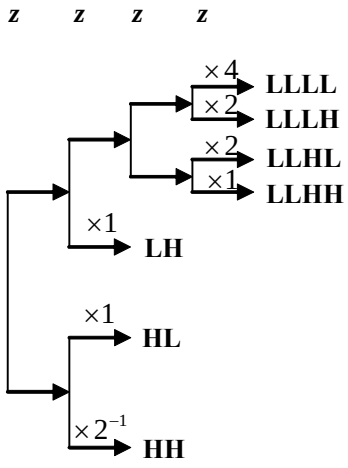

FIG. 1.10: Structure en paquet d'ondelettes 1D sur 4 niveaux permettant de rendre la transformation approximativement unitaire par un décalage implicite des coefficients d'ondelettes entiers.

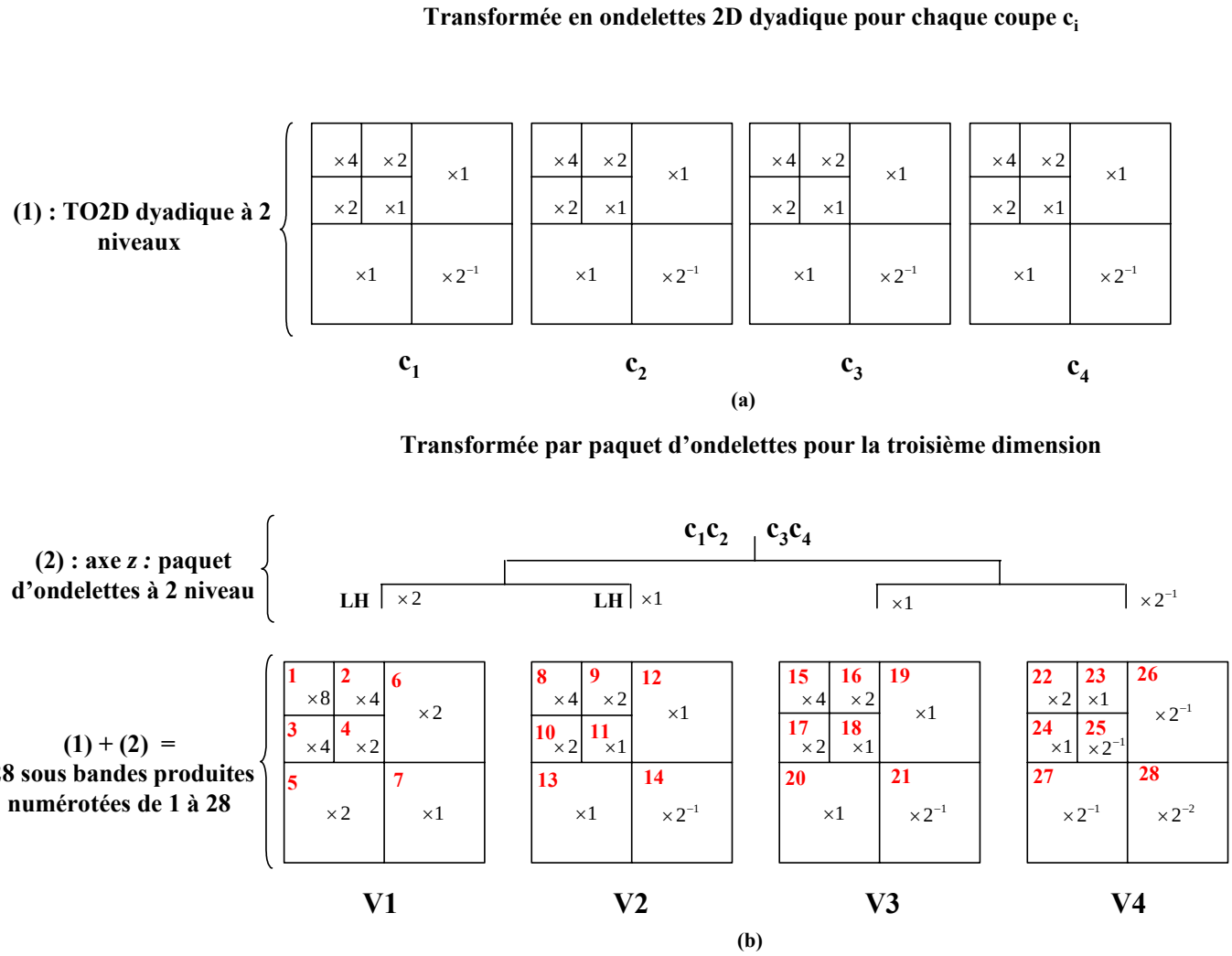

Fig. 1.11: Facteurs à utiliser pour un décalage implicite des coefficients d'ondelettes entiers pour approximer une TO3D unitaire avec un lifting 5.3.

figure 1.11 (a). Nous n'obtenons bien que des puissances de 2. Enfin, puisque l'implantation des facteurs de normalisation se fait à travers un décalage de bit, les facteurs de normalisation doivent être en pratique multipliés par le plus petit entier qui est une puissance de 2 pour que les facteurs de normalisation ne soient pas des puissances de 2 négatives. Sinon, des bits serait perdus par des décalages par la droite pour les facteurs de normalisation fractionnaires. Dans l'exemple donné dans la figure 1.11 (b), tous les facteurs seraient multipliés par quatre.

1.4.5 Discussion à propos des deux techniques de normalisation

La normalisation par décalage de bits grâce à une structure par paquets d'ondelettes est plus simple à implanter qu'une par lifting (sans prendre en compte les différences entre la transformée dyadique et par paquet d'ondelettes). En effet, elle ne nécessite pas d'étape supplémentaire de lifting. Cependant, elle est moins flexible car les facteurs de normalisation doivent être $\sqrt{2}$ pour les coefficients passe-bas et $\frac{1}{\sqrt{2}}$ pour les coefficients passe-haut. Par exemple, la TO3D avec un filtre 9/7 F entier ne peut pas être implantée par paquet d'ondelettes car les facteurs de normalisation ne sont pas $\sqrt{2}$ et $\frac{1}{\sqrt{2}}$. Une autre limitation pour utiliser le décalage de bits pour la normalisation

est que le nombre de décompositions dans la direction temporelle (axe z) doit être obligatoirement pair.

Au final, insistons sur le fait que la TO3D par paquet d'ondelettes est radicalement différente d'une TO3D dyadique classique. Dans la figure 1.11, notons que 28 sous-bandes (numérotées de 1 à 28 sur la figure) ont été produites. C'est le double de sous-bandes qu'une TO3D dyadique aurait produit. Cela est dû à la transformée supplémentaire en z liée à la structure par paquet d'ondelettes dans cette direction. La référence [58] montre la décomposition en arbre de la TO3D par paquets d'ondelettes de la figure 1.11. Notons enfin qu'une TO3D par paquet d'ondelettes donnera des performances très proches d'une TO3D dyadique pour le codage sans perte, même quand aucun facteur de normalisation n'est employé pour les deux [118]. Ainsi, appliquer la même TO2D sur chaque coupe de l'image 3D suivie d'une TO1D dans la troisième dimension exploitera aussi efficacement la corrélation parmi les coupes voisines d'une image médicale volumique qu'une transformée dyadique.

1.5 Méthodes actuelles de codage des images médicales.

Après la première étape de transformée permettant la décorrélation des données, les coefficients produits entiers ou réels suivant l'implantation de la TO3D peuvent être codés avec ou sans perte. Nous distinguons deux types d'approches pour les standards de compression actuels basés sur la TO : l'approche inter-bandes qui utilise les redondances inter-échelles entre les sous-bandes pour coder les coefficients d'ondelettes et celle intra-bandes dans laquelle les sous-bandes sont codées de façon indépendante.

1.5.1 Méthodes inter-bandes

Le premier algorithme inter-bandes pour les images 2D se nomme EZW[9][104] et a été proposé en 1993 par Jerry Shapiro. L'ensemble des méthodes qui ont suivi s'appuient sur des techniques communes. Elles exploitent complètement la notion de multi-résolution associée aux ondelettes. Leur schéma de codage utilise un modèle simple pour caractériser les dépendances inter-bandes parmi les coefficients d'ondelettes localisés dans les sous-bandes ayant la même orientation. La figure 1.5 illustre ces dépendances à travers toutes les échelles. Le modèle est basé sur l'hypothèse des arbres de zéros, qui suppose que si un coefficient d'ondelettes w est non significatif pour un seuil donné T, c'est à dire $|w| < T$, alors tous les coefficients de la même orientation dans la même localisation spatiale à des résolutions plus fines sont supposés non significatifs pour ce même seuil T. Cette hypothèse des arbres de zéros est illustrée en 2D sur la figure 1.12. On peut voir qu'un coefficient "parent" à une résolution donnée va engendrer 4 coefficients "enfant" à la résolution supérieure.

Ces méthodes appliquent une quantification par approximations successives pour améliorer la précision de la représentation des coefficients d'ondelettes et pour faciliter le codage imbriqué. Avec cette approche, la signification des coefficients d'ondelettes pour une série de seuils monotone décroissante T_n est enregistrée dans un ensemble de cartes binaires, appelées cartes de signification correspondant chacune à un plan de bits. La figure 1.13 illustre cette notion de plan de bits. Par ailleurs, Shapiro a prouvé que le codage des cartes de signification représente une part importante du coût d'encodage total. Ainsi, en améliorant l'encodage de ces cartes de signification, on peut s'attendre à un gain de codage significatif. La technique utilisée pour encoder les cartes de signification est un algorithme par arbre de zéros, qui permet un codage peu coûteux en terme de débit.

Par ailleurs, ce type de méthodes permet de faire de la transmission progressive (amélioration de la qualité par ajout de bits dans ce cas). En effet, elles peuvent trier l'ordre des bits de codage

[9] *Embedded image coding using Zerotrees of Wavelet coefficients.*

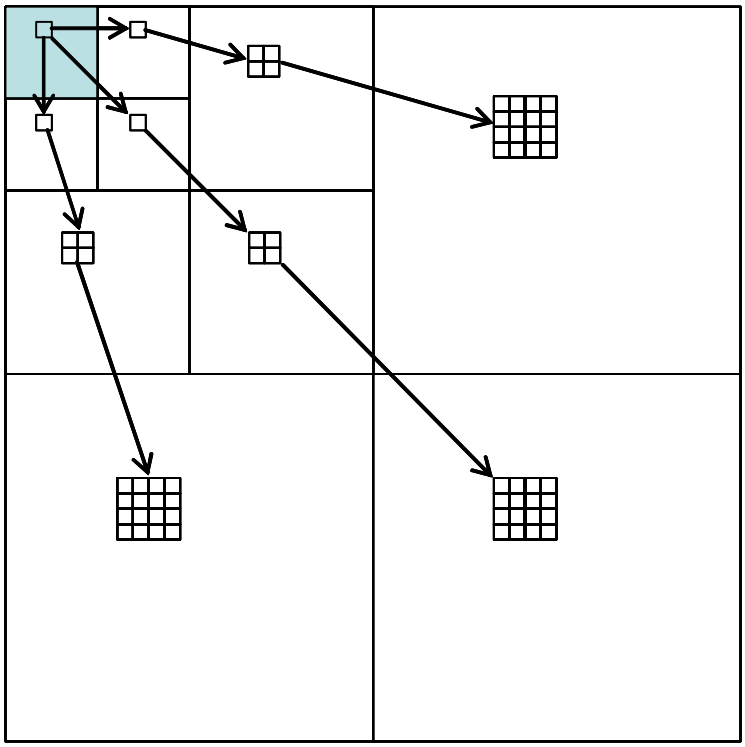

Fig. 1.12: Illustration de l'approche inter-bandes : structure arborescente de SPIHT et de EZW dans le cas 2D.

tel que les bits les plus significatifs soient transmis en premier. Ainsi, pour augmenter la qualité de l'image reconstruite, il suffira d'ajouter de la précision en utilisant les bits supplémentaires.

1.5.1.1 EZW 3D : extension 3D de EZW

EZW 3D [10] [70] est l'extension aux images 3D volumiques d'EZW qui s'appuie sur des relations parents/enfants dans trois dimensions au lieu de deux. Ceci est lié à l'utilisation d'une TO3D dyadique présentée précédemment. La figure 1.14 illustre les nouvelles dépendances.

Notons que le noeud racine (correspondant aux coefficients de la sous-bande LLL_D pour une TO3D à D niveaux de décomposition) de l'arbre a seulement 7 enfants, alors que tous les autres noeuds à l'exception des extrémités en possèdent 8. En d'autres termes, à l'exception du noeud racine et des extrémités de l'arbre, le lien parent enfant pour EZW 3D est le suivant :

$$O(i,j,k) = \left\{ \begin{array}{c} (2i,2j,2k),(2i,2j+1,2k),(2i+1,2j,2k), \\ (2i+1,2j+1,2k),(2i,2j,2k+1),(2i+1,2j,2k+1), \\ (2i,2j+1,2k+1),(2i+1,2j+1,2k+1) \end{array} \right\} \tag{1.44}$$

où $O(i,j,k)$ représente un ensemble de coordonnées de tous les enfants du noeud (i,j,k). Pour définir la relation parent enfant pour un coefficient du noeud racine, considérons l_x, l_y et l_z les dimensions de la sous-bande racine (sous-bande basse fréquence de la TO3D). Un coefficient de la sous-bande racine (LLL_D) de coordonnées (i,j,k) possède comme enfants les coefficients suivants :

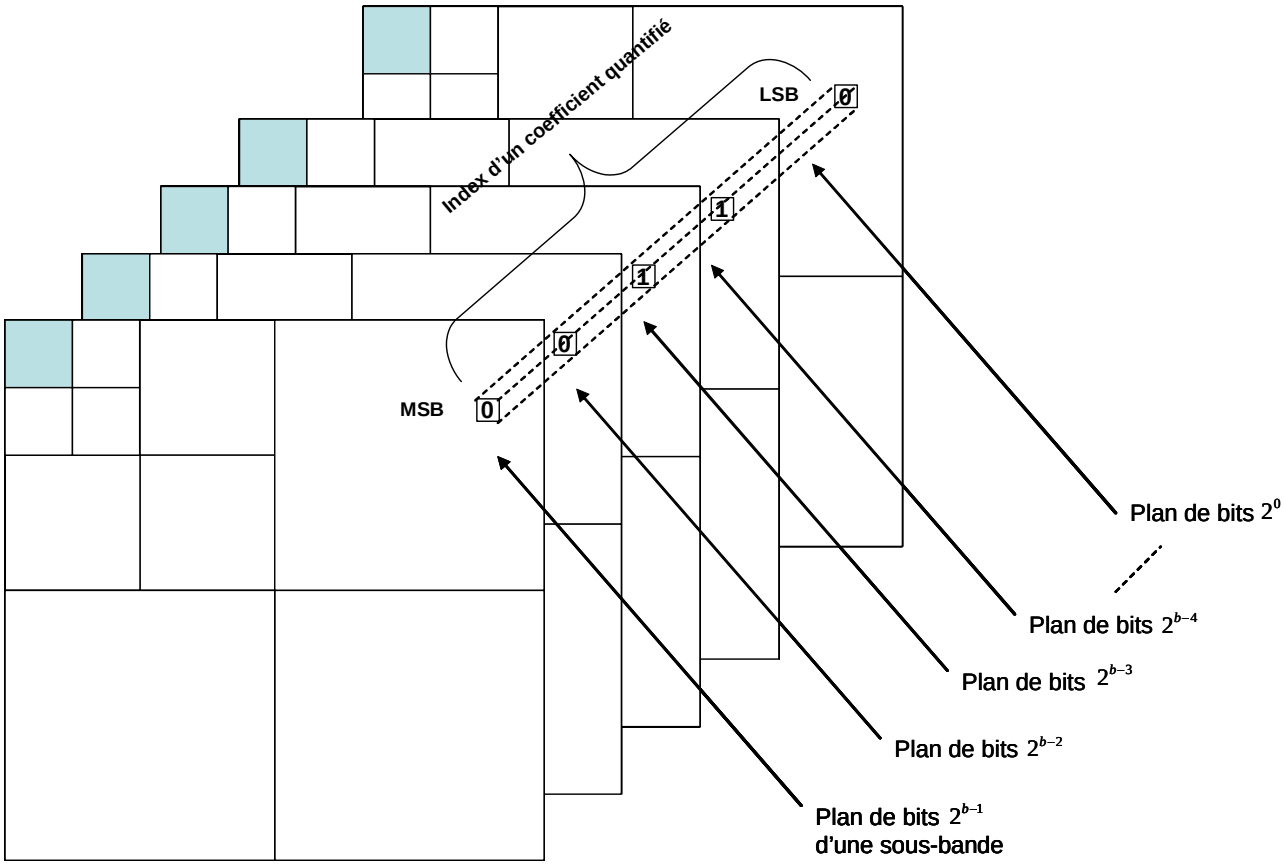

FIG. 1.13: Représentation des plans de bits allant du bits le plus significatif (MSB) vers le moins significatif (LSB) pour une transformée en ondelettes 2D.

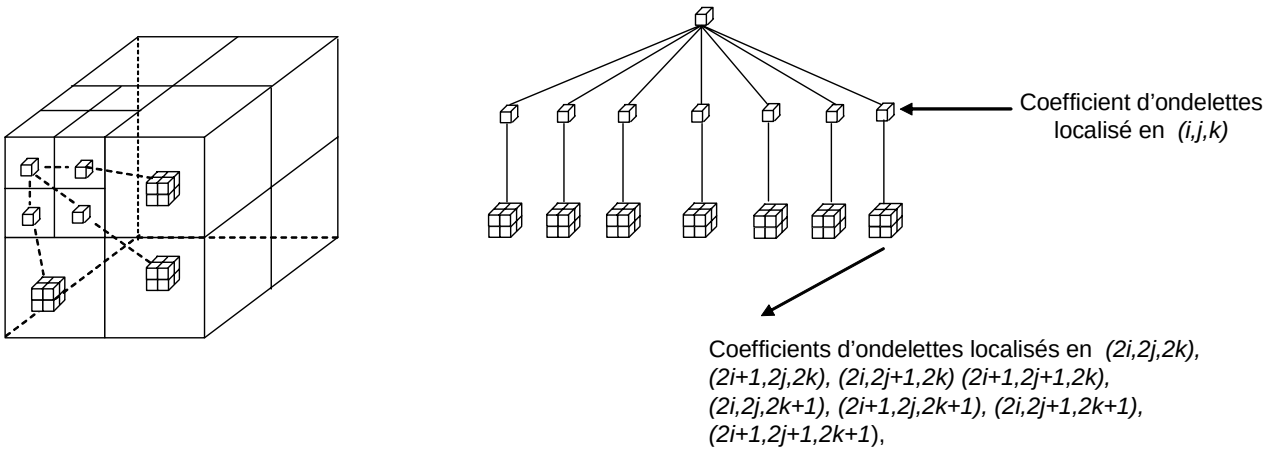

FIG. 1.14: Diagramme de la structure en arbre 3D de EZW 3D.

$$O_{LLL}(i,j,k) = \left\{ \begin{array}{c} (i+l_x,j,k),(i+,j+l_y,k),(i,j,k+l_z), \\ (i+l_x,j+l_y,k),(i+l_x,j,k+l_z), \\ (i,j+l_y,k+l_z),(i+l_x,j+l_y,k+l_z) \end{array} \right\} \tag{1.45}$$

La suite de l'algorithme ressemble au cas 2D. Ainsi, on part d'un seuil entier T_n qui est une puissance de 2 de type $T_n = 2^{\lfloor \log_2(w_{\max}) \rfloor} = 2^n$ où $w_{\max}$ est le coefficient le plus grand en valeur absolue de toutes les sous-bandes. Les coefficients d'ondelettes de la TO3D sont scannés dans un ordre hiérarchique de la résolution la plus faible vers la plus grande, et chaque coefficient w est testé pour savoir si sa valeur absolue est supérieure ou égale au seuil T_n , c'est à dire si il est significatif pour le seuil donné. Si un coefficient est trouvé significatif, il est codé suivant son signe par le symbole POS pour un coefficient positif et NEG pour un coefficient négatif. Il est alors placé dans la liste des coefficients significatifs.

Si un coefficient est testé non significatif ($|w| < T_n$), on examine tous ses descendants pour tester leur signification. Dans le cas où aucun descendant n'est significatif, on code un arbre de zéros (AZ). Si un descendant significatif apparaît on code un zéro isolé (ZI). Les coefficients qui descendent d'une racine d'arbre de zéros n'ont pas besoin d'être codés. Cette partie se nomme la passe dominante pour Shapiro ou passe de signification dans de nombreux articles.

Ensuite, la seconde passe (passe de raffinement ou subalterne pour Shapiro) est réalisée sur les coefficients dans la liste significative. Pour chaque coefficient de cette liste, le bit situé dans le plan de bit inférieur (plan de bit 2^{n-1}) est codé. L'encodeur divise le seuil T_n par 2 ($T_n \leftarrow \frac{T_n}{2}$) et exécute une nouvelle passe dominante et de raffinement. Cette procédure se poursuit jusqu'à ce que l'on atteigne le débit voulu. Si un coefficient est trouvé significatif à une passe postérieure, il sera encore dans la liste significative à la passe courante et n'aura pas besoin d'être identifié comme significatif une autre fois. Le nombre de coefficients dans cette liste croît de façon monotone au fur et à mesure que les seuils T_n décroissent. Si l'on va jusqu'au dernier plan de bits (LSB), on obtient un codage sans perte car il n'y a plus d'étape de quantification.

Le décodeur utilise un algorithme similaire. Il initialise tous les coefficients à zéro et scanne à travers les mêmes directions que l'encodeur. Le décodeur reçoit un symbole du flux binaire pour chaque coefficient. Si ce symbole est POS ou NEG, l'amplitude du coefficient est au dessus du seuil et on détermine le signe. Dans les deux cas, le coefficient est placé dans la liste significative. Si un symbole AZ est reçu, aucun des descendants du coefficient courant n'est visité pendant la passe dominante.

On réalise ensuite la passe de raffinement. Pour chaque coefficient dans la liste de signification, un bit est sorti du flux binaire. Si c'est un 1, il est utilisé pour remplacé le bit 0 à la localisation $\log_2(T_n) - 1$ dans la représentation binaire des coefficients.

Revenons à présent au flux binaire émis par le codeur. Afin d'optimiser les performances débit/distorsion, celui-ci doit être codé de manière entropique. Pour cette approche, le meilleur type de code entropique est un code arithmétique basé sur contexte car il exploite très efficacement la non uniformité des probabilités et les dépendances entre les symboles [115]. La version d'EZW 3D où le codage des symboles s'effectue par un codage arithmétique basé sur contexte se nommera CB EZW 3D. A contrario, la version simple d'EZW 3D utilisera quant à elle un simple codeur arithmétique.

Le point fondamental de la conception codeur arithmétique adaptatif basé sur contexte est le choix d'un bon modèle de contexte. Si l'on considère w le symbole ou coefficient d'ondelette que l'on veut coder, un simple modèle sans mémoire ne peut pas être efficace et nécessiterait $-\log_2(P(w))$ bits pour encoder ce symbole. Un meilleur choix est d'utiliser un modèle d'ordre supérieur. Soit un modèle de contexte, $C = \{w_1, w_2, ..., w_L\}$, où les w_i sont les symboles des autres coefficients qui dépendent du symbole courant. Ainsi $-\log_2(P(w|C))$ bits sont nécessaires pour encoder w. Notons que si chaque w_i possède B bits de longueur, il y a 2^{BL} contextes différents.

En théorie, plus l'ordre du modèle de contexte augmente, plus l'entropie conditionnelle diminue

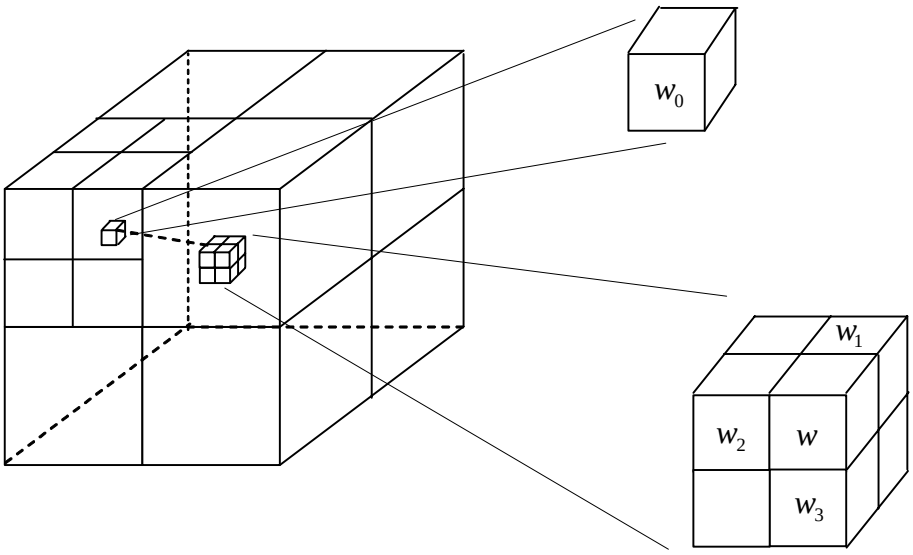

FIG. 1.15: Diagramme du modèle de contexte 3D de EZW 3D pour le symbole courant w.

[31]. Cependant, dans la pratique, augmenter l'ordre du modèle de contexte n'améliore pas toujours la performance de codage. Le codeur arithmétique nécessite une estimation du modèle statistique de la source $P(w|C)$ qui doit être évalué à la volée à partir des observations passées. Etant donné que la quantité de données nécessaires pour estimer $P(w|C)$ croît avec l'augmentation de l'ordre du modèle, un modèle de contexte d'un très grand ordre peut entraîner de nombreux symboles codés par l'utilisation d'estimations de probabilités inadaptées. Ce problème est connu sous le problème de dilution de contexte [95].

Pour le codage arithmétique adaptatif basé sur contexte des symboles dans EZW 3D, les modèles de contexte peuvent aussi bien utiliser les dépendances hiérarchiques et spatiales que les dépendances à l'intérieur des sous-bandes d'un même niveau de la transformée. Ainsi, les coefficients d'ondelettes autour du coefficient courant peuvent être utilisés pour exploiter les dépendances spatiales et le coefficient parent peut être utilisé pour prendre avantage des dépendances hiérarchiques.

De façon similaire, les dépendances entre les sous-bandes d'un même niveau de résolution peuvent être exploitées en utilisant les coefficients à la même localisation spatiale de l'image. Cependant, il est fondamental de préserver la causalité des modèles de contexte. En effet, les modèles de probabilité utilisés pour exécuter le codeur arithmétique sont sélectionnés sur la base du contexte de chaque symbole. Si le contexte sélectionné diffère pour un symbole entre l'encodeur et le décodeur, le décodeur sélectionnera une table de probabilité différente. Par exemple, si l'information d'un coefficient d'ondelette qui est non causal par rapport à la direction de parcours est utilisée pour former un contexte à l'encodeur, le décodeur ne sera pas capable de dupliquer ce contexte car cette information non causale n'est pas encore disponible pour le décodeur. Cela entraînera une perte de synchronisation et le reste du flux binaire sera perdu. Ainsi, chaque contexte doit être formé uniquement par l'information qui est dèjà disponible pour le décodeur.

Dans la majorité des études qui ont implanté EZW 3D ([10], [69], [70]), le modèle de contexte 3D mixe l'utilisation des dépendances spatiales et hiérarchiques. Par exemple, dans [10], on utilise à la fois trois symboles autour du symbole à encoder et le symbole parent pour créer le modèle de contexte. Ce modèle est illustré sur la figure 1.15. Les symboles w_1, w_2 et w_3 autour du symbole courant w peuvent prendre les 4 valeurs : POS, NEG, ZI, AZ tandis que le parent w_0 est seulement identifié comme significatif ou non significatif. Ce schéma produits 128 contextes. Notons que les 4 symboles w_0, w_1, w_2 et w_3 sont choisis tels qu'ils soient disponibles quand le symbole w est décodé.

La complexité de EZW 3D basé sur contexte (CB EZW 3D) est comparable à celle de EZW 3D, c'est à dire très faible. Le seul ajout de complexité est lié au calcul de l'index de la table de probabilité auquel appartient le contexte courant. Les besoins en mémoire de CB EZW 3D sont légèrement supérieurs à celle sans contexte. Une simple table de probabilité est suffisante pour un simple codage arithmétique adaptatif des symboles de EZW 3D alors que CB-EZW 3D requiert une table de probabilité par contexte. Ainsi, 128 tables de probabilités ont besoin d'être gardées en mémoire pour un modèle avec 128 contextes. Chaque modèle de probabilité requiert un stockage de la fréquence d'apparitions de chaque symbole. Par exemple, si l'on prend 2 octets pour stocker chaque fréquence d'apparition, l'algorithme CB-EZW 3D nécessite 2 octets $\times$ 4 symboles $\times$ 128 contextes = 1024 octets pour stocker les tables de probabilité tandis que EZW 3D ne requiert que 2 octets $\times$ 4 symboles = 8 octets.

Son successeur SPIHT, plus élaboré va encore améliorer les performances de codage en repoussant le compromis débit-distorsion sans augmenter la complexité.

1.5.2 SPIHT 3D

1.5.2.1 Principe

L'algorithme SPIHT[10] a été proposé par Saïd et Pearlman en 1996 pour la compression avec [98] et sans perte [99]. Il a été étendu au 3D pour la vidéo [57] et pour la compression d'images volumiques [58]. Cet algorithme repose sur la même idée que celle de Shapiro (EZW) pour caractériser les dépendances entre les coefficients d'ondelettes. Cependant, il est à la fois plus complexe et plus efficace pour coder les cartes de signification.

Les principes de base de SPIHT 3D sont identiques à sa version 2D : un rangement partiel par amplitude des coefficients d'ondelettes de la TO3D (résultant de la quantification par approximations successives), un partitionnement dans des arbres hiérarchiques (à chaque seuil appliqué les arbres sont triés sur la base de leur signification en deux catégories d'arbre et un ordonnancement de la transmission des bits de raffinement (l'amplitude de chaque coefficient significatif est progressivement raffinée). Sa première implantation est basée sur des arbres à orientation spatio-temporelle équilibrée [57]. Par conséquent, le même nombre de décompositions récursives d'ondelettes est requis pour les trois directions spatiales (x, y et z). Si cela n'est pas respecté, plusieurs nœuds de l'arbre ne sont pas rattachés ou sont liés avec la même localisation spatiale et par conséquent les dépendances entre les nœuds de l'arbre sont détruites et ainsi la performance de compression. On peut également mentionner que des solutions à ce problème ont été proposées en utilisant des arbres à orientation spatio-temporelle non équilibrée (AT - SPIHT 3D) dans [19]. Cela permet d'appliquer facilement un nombre différent de décompositions entre les dimensions spatiales et temporelles, qui est une fonctionnalité souhaitable quand on code un ensemble de coupes qui est limité en taille.

La différence essentielle entre EZW 3D et SPIHT 3D est la façon dont les coefficients des arbres sont construits, triés et découpés. Ainsi la structure même des arbres de zéros est différente. Dans EZW 3D, un arbre de zéros est défini par un coefficient racine et ses descendants ont tous la valeur zéro à l'intérieur d'un plan de bits. SPIHT 3D utilise lui deux types d'arbres de zéros. Le premier (type A) consiste en une simple racine ayant tous ses descendants à 0 pour un plan de bits donné. Cela diffère un peu des arbres de zéros de EZW 3D du fait que la racine elle-même n'a pas besoin d'être non significative. En fait, bien que l'arbre de zéros soit spécifié par les coordonnées de la racine, la racine n'est pas incluse dans l'arbre. Le second type d'arbre (type B) est similaire mais exclut les huit enfants de la racine. Les arbres de type B contiennent uniquement les petits-enfants, arrières petits-enfants ... de la racine. De plus, dans SPIHT 3D, les arbres sont définis de tel façon que chaque noeud ne possède aucun descendant (les feuilles) ou bien 8 descendants qui forment un groupe adjacent de $2\times 2\times 2$. Les coefficients de la sous-bande basse fréquence LLL_D correspondant

[10] *Set Partitioning In Hierarchical Trees*

aux racines de l'arbre sont également groupés en coefficients $2 \times 2 \times 2$ adjacents. Cependant, la relation parent enfant pour un coefficient du noeud racine est altérée par rapport à EZW 3D. Dans chaque groupe de $2 \times 2 \times 2$ de LLL_D un des coefficients n'a pas de descendants. Ainsi, tous les coefficients $w(i,j,k)$ qui possèdent trois coordonnés impaires (i,j,k) n'ont pas de descendant. En fait, la relation O_{LLL} (1.45) n'est plus valable dans SPIHT 3D pour caractériser les relations parent enfant dans cette sous bande.

Les ensembles suivants de coordonnées sont utilisés dans la méthode de codage complète présenté par la suite :

- $O(i,j,k)$: Ensemble des coordonnées de tous les enfants du noeud (i,j,k). Il s'exprime de la même façon que celui de EZW 3D
- $D(i,j,k)$: Ensemble des cordonnées de tous les descendants du noeud (i,j,k) (type A d'arbres de zéros)
- $L(i,j,k) = D(i,j,k) - O(i,j,k)$ (type B d'arbre de zéros)

Les règles de partitions sont les suivantes :

1. La partition intiale est formée des ensembles $\{(i,j,k)\}$ et $D(i,j,k)$, pour tous $(i,j,k) \in LLL_D$ qui ont un descendant.
2. Si $D(i,j,k)$ est significatif alors il est découpé en $L(i,j,k)$ plus 8 ensembles d'un seul élément avec $(l,m,n) \in O(i,j,k)$.
3. Si $L(i,j,k)$ est significatif alors il est partitionné en 8 sous-ensembles $D(l,m,n)$ avec $(l,m,n) \in O(i,j,k)$.

1.5.2.2 Algorithme de codage

Pour réaliser pratiquement son codage imbriqué, SPIHT 3D stocke l'information de signification dans 3 listes ordonnées :

- la Liste des Coefficients Significatifs (LCS),
- la Liste des Coefficients Non significatifs (LCN),
- la Liste des Ensembles Non significatifs (LEN).

Dans chaque liste, l'entrée est représentée par un triplet de coordonnées (i,j,k), qui représentent dans LCS et LCN des coefficients individuels et dans LEN soit l'ensemble $D(i,j,k)$ soit $L(i,j,k)$.

Pendant la passe de signification, les coefficients dans LCN, qui étaient non significatifs dans la passe précédente sont testés. Ceux qui deviennent significatifs sont mis dans LCS. Similairement, les ensembles de LEN sont évalués dans leur ordre d'entrée, et quand un ensemble est trouvé significatif il est supprimé de cette liste puis est partitionné. Les nouveaux ensembles avec plus d'un élément sont ajoutés à la fin de LEN avec le type (A ou B), alors que les simples coefficients sont ajoutés à la fin de LCS ou LCN suivant leur signification. La liste LCS contient les coordonnées des coefficients qui seront visités dans la prochaine passe de raffinement.

Nous rappelons l'algorithme de base de SPIHT 3D [57]. Pour cela, nous définissons l'opérateur de signification σ_{T_n} qui évalue la signification d'un sous-ensemble E pour un seuil donné T :

$$\sigma_{T_n}(E) = \begin{cases} 1 & \text{si } \exists w \in E : |w| \geq T_n \\ 0 & \text{si } \forall w \in E : |w| < T_n \end{cases} \tag{1.46}$$

1. **Initialisation** : *Sortie* $n = \lfloor \log_2(w_{\max}) \rfloor \Leftrightarrow T_n = 2^{\lfloor \log_2(w_{\max}) \rfloor}$; $LCS = \emptyset$; $LCN = \{(i,j,k) \in LLL_D\}$. LEN contient les mêmes coefficients que LCN excepté ceux qui n'ont pas de descendants.
2. **Passe de signification**

   2.1 Pour chaque $(i,j,k) \in LCN$ faire :

2.1.1 *Sortie* $\sigma_{T_n}(i,j,k)$

2.1.2 si $\sigma_{T_n}(i,j,k)=1$ alors mettre (i,j,k) dans LCS et coder le signe de $w(i,j,k)$

2.2 Pour chaque $(i,j,k) \in LEN$ faire :

2.2.1 Si l'entrée est de type A

a *Sortie* $\sigma_{T_n}(D(i,j,k))$

b si $\sigma_{T_n}(D(i,j,k))=1$ alors

- Pour chaque $(l,m,n) \in O(i,j,k)$ faire :
  - *Sortie* $\sigma_{T_n}(l,m,n)$
  - si $\sigma_{T_n}(l,m,n)=1$ alors mettre (l,m,n) dans LCS et coder le signe de $w(l,m,n)$
  - si $\sigma_{T_n}(l,m,n)=0$ alors mettre (l,m,n) à la fin de LCN
- Si $L(i,j,k) \neq \emptyset$ alors mettre (l,m,n) à la fin de LEN comme une entrée de type B

2.2.2 Si l'entrée est de type B

a *Sortie* $\sigma_{T_n}(L(i,j,k))$

b si $\sigma_{T_n}(L(i,j,k))=1$ alors

- mettre (l,m,n) à la fin de LEN comme une entrée de type A
- supprimer (i,j,k) de LEN

3 **Passe de raffinement** : Pour chaque coefficient $(i,j,k) \in LCS$ à l'exception de ceux incluse dans la même passe de signification (c.a.d pour le même n). *Sortie* le n-ième bit significatif de $|w(i,j,k)|$

4 **Modification du pas de quantification** : $T_n \leftarrow \frac{T_n}{2}$ et aller à l'étape **2**

Nous remarquons par rapport à EZW 3D l'inversion des passes de raffinement et de signification. Ainsi, chaque plan de bits est codé par une passe de signification suivi d'une passe de raffinement alors que dans EZW 3D, la passe de raffinement code un bit de raffinement pour chaque coefficient qui était significatif à la fin du plan de bits précédent. Ainsi, les coefficients qui deviennent significatifs via la passe de signification du plan de bits courant ne sont pas raffinés jusqu'au prochain plan de bits dans SPIHT 3D.

Pour obtenir, l'algorithme de décodage, il suffit simplement de remplacer le mot *Sortie* par *Entrée* dans l'algorithme précédent.

De plus, le décodeur exécute une tâche supplémentaire en modifiant l'image reconstruite. Pour un seuil T_n donné, quand un coefficient est déplacé dans la LCS, il est évident que $T_n \leq w(i,j,k) < 2T_n = 2^{n+1}$. Ainsi, le décodeur utilise cette information plus le bit de signe juste après l'insertion dans la LCS pour mettre $\widehat{w}(i,j,k) = \pm 1.5 \times T_n$. De manière identique, pendant la passe de raffinement le décodeur ajoute ou soustrait $\frac{T_n}{2}$ à $\widehat{w}(i,j,k)$ quand on reçoit les bits de la représentation binaire de $|w(i,j,k)|$. De cette manière, la distorsion baisse à la fois pendant les 2 passes.

Enfin, on notera que contrairement à EZW 3D, SPIHT 3D produit directement des symboles binaires. Ainsi, un codeur arithmétique basé sur contexte n'est pas obligatoire même si il est souvent implanté pour améliorer les performances de codage. Les modèles de contexte sélectionnés sont basés sur la signification des nœuds individuels, aussi bien que l'état de leurs descendants. Par conséquent, pour chaque coefficient de nœud, 4 combinaisons d'états sont possibles. Au total, un modèle avec 164 contextes est utilisé pour le codage arithmétique pendant la passe de signification dans notre version de SPIHT 3D. Par ailleurs, le codeur arithmétique n'est pas implanté dans le codage du signe et lors de la passe de raffinement car le gain entropique est négligeable.

L'ensemble des améliorations proposées dans SPIHT 3D par rapport à EZW 3D en fait la référence des méthodes de codage inter-bandes.

1.5.3 Méthodes intra-bandes

Au contraire des méthodes inter-bandes comme EZW ou SPIHT, dans les méthodes inter-bandes, les sous-bandes produites par la TO3D sont codées de manière indépendante. Au sein de cette famille, on distinguera deux types d'approches. La première s'appuie sur une logique par plan de bits assez semblable aux méthodes inter-bandes mais n'utilise pas d'approche par arbres de zéros. Typiquement, le codage de chaque plan de bits comprend le codage de la signification, le codage du signe et le raffinement de l'amplitude. Les quatre premiers codeurs présentés sont de cette nature et correspondent aux méthodes intra-bandes de référence en imagerie médicale. La seconde approche diffère en ce sens qu'elle est basée sur une procédure d'optimisation qui alloue un nombre de bits à priori pour chaque sous-bande tout en minimisant l'erreur totale. Nous proposons dans ce cadre une nouvelle méthode basée sur un quantificateur vectoriel algébrique, que nous présenterons dans le chapitre 2. Au contraire de toutes les autres méthodes présentées dans ce paragraphe qui sont basées sur des codeurs performants par plan de bits, **notre approche est davantage centrée sur une quantification optimale que sur le codage.**

1.5.3.1 Le Cube Splitting (CS)

L'algorithme Cube splitting [101] est dérivé du codeur SQP[11] de Munteanu et Cornelis. SQP est basé sur une quantification par approximations successives et utilise un codage en arbre quaternaire des cartes de signification. Comme dans l'approche inter-bandes, chaque plan de bits est encodé en deux étapes classiques : une passe de signification et une autre de raffinement. Cependant, ce codeur n'emploie pas la même décomposition des cartes de signification que la famille des codeurs basés sur les arbres de zéros. En effet, les sous-bandes sont codées de manière indépendante car le partitionnement isolera des sous-bandes (volumes) non significatives pour un plan de bits donné qui ne seront pas codées.

On part encore une fois de $T_n = 2^{\lfloor \log_2(w_{\max}) \rfloor}$ le seuil maximum utilisé pour la quantification par approximations successives des coefficients d'ondelettes. On note $V(\mathbf{c}, \mathbf{l})$ un volume de coefficients qui a pour coordonnées en haut à gauche $\mathbf{c} = (i, j, k)$ et pour taille $\mathbf{l} = (l_x, l_y, l_z)$, où l_x, l_y et l_z sont respectivement la longueur, la largeur et la profondeur du volume. Pour simplifier, nous supposons que les dimensions sont toutes des puissances de 2. Finalement, nous considérerons l'image volumique d'ondelettes $\mathbf{W}$ produites par la TO3D comme un tenseur de $L_x \times L_y \times L_z$ éléments. L'image volumique d'ondelettes est donc définie comme $\mathbf{W} = V(\mathbf{0}, \mathbf{L})$, avec $\mathbf{0} = (0, 0, 0)$ et $\mathbf{L} = (L_x, L_y, L_z)$. La signification d'un volume $V(\mathbf{c}, \mathbf{l})$ pour un seuil donné T est déterminée par l'opérateur de signification σ_{T_n} :

$$\sigma_{T_n}\left(V(\mathbf{c}, \mathbf{l})\right) = \begin{cases} 1 & \text{si } \exists w \in V(\mathbf{c}, \mathbf{l}) : |w| \geq T_n \\ 0 & \text{si } \forall w \in V(\mathbf{c}, \mathbf{l}) : |w| < T_n \end{cases} \tag{1.47}$$

Avec w un coefficient d'ondelettes du volume V. On peut remarquer que pour $\mathbf{l} = (1, 1, 1)$, l'opérateur de signification σ_{T_n} déterminera la signification d'un simple coefficient d'ondelettes w.

On applique l'algorithme CS sur le volume complet d'ondelettes produit par la TO3D pour isoler des petites entités, c'est-à-dire des sous-cubes qui contiennent potentiellement des coefficients significatifs d'ondelettes. Pendant la première passe de signification, on teste pour le plus grand plan de bits (seuil T_n), la signification de l'image d'ondelettes (volume) $\mathbf{W}$ avec l'opérateur de signification σ_{T_n} défini précédemment. Si $\sigma_{T_n} = 1$, l'image d'ondelettes est partitionnée en 8 sous

[11] *SQuare Partitioning coding.*

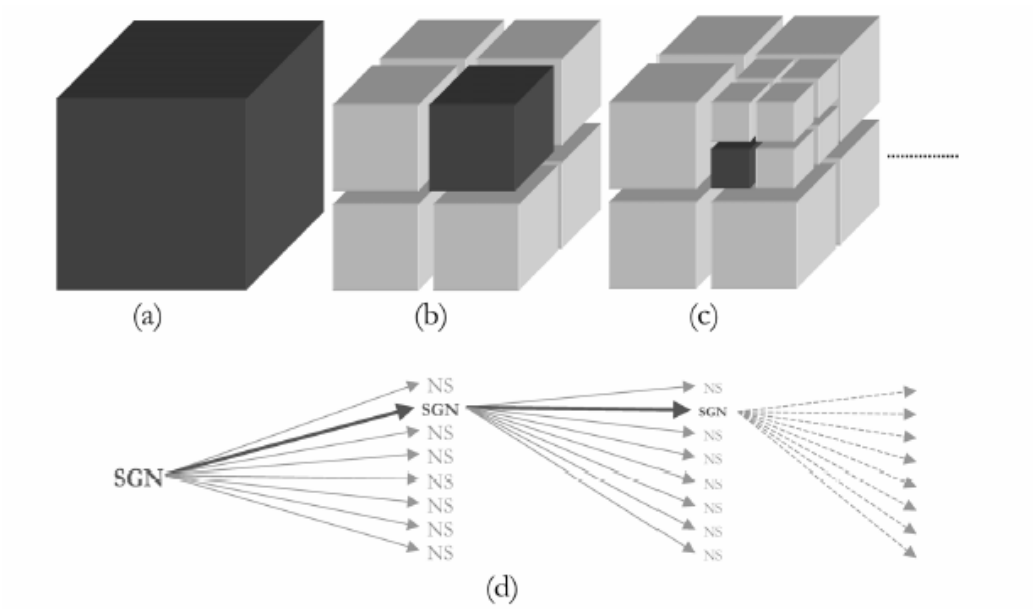

Fig. 1.16: Illustration de la partition effectuée par le Cube Splitting en isolant les coefficients significatifs. Quand un coefficient d'ondelettes significatif est rencontré, le cube (a) est découpé en huit sous-cubes (b), et ainsi de suite (c) jusqu'au niveau du pixel. Le résultat est une structure en arbre en octons. (d) (SGN = noeud significatif ; NS = noeud non significatif).

volumes : $V\left(\mathbf{c}_m, \frac{\mathbf{l}}{2}\right), 1 \leq m \leq 8$, avec des coordonnées en haut à gauche $\mathbf{c}_m = (i, j, k)$ et une taille diminuée de moitié dans les 3 dimensions de $\frac{\mathbf{l}}{2} = \left(\frac{l_x}{2}, \frac{l_y}{2}, \frac{l_z}{2}\right)$.

Le ou les cubes significatifs descendants sont découpés jusqu'à ce que les coefficients d'ondelettes significatifs $\sigma_{T_n}(w) = 1$ soient isolés. Ainsi, la phase de signification marque des sous-cubes et des coefficients d'ondelettes, nouvellement identifiés comme significatifs, en utilisant une structure en arbre récursive d'octons. Le résultat est un schéma structuré en octons pour la signification des données par rapport à un seuil donné. Comme, on peut le noter des poids égaux sont donnés sur toutes les branches. Quand un coefficient significatif est isolé, son signe pour lequel 2 symboles de code sont préservés est immédiatement codé. La figure 1.16 illustre cette structure en octons.

Quand le plan de bit complet est codé par cette passe de signification, l'encodeur divise le seuil T_n par 2 ($T_n \leftarrow \frac{T_n}{2}$). Puis, la passe de raffinement est initiée en raffinant tous les coefficients marqués comme significatifs dans l'arbre.

La passe de signification est ensuite reproduite pour modifier l'arbre en identifiant les nouveaux coefficients d'ondelettes significatifs pour le plan de bit courant. Pendant cette étape, seuls les nœuds non significatifs précédemment sont traités pour signification et ceux qui étaient significatifs à l'étape précédente sont ignorés puisque le décodeur a déjà reçu l'information. La procédure décrite est répétée jusqu'à ce que l'image complète d'ondelettes soit encodée, c'est-à-dire jusqu'à ce qu'on ait atteint le dernier plan de bits ou que l'on ait obtenu le débit désiré.

Pour encoder efficacement, les symboles générés, un encodeur arithmétique basé sur contexte est intégré. Le modèle de contexte est simple. Pour la passe de signification, quatre modèles de contexte sont distingués, typiquement un pour les symboles générés au nœud du cube intermédiaire, un pour les nœuds de coefficient ayant des voisins non significatifs pour le seuil précédent, un pour les nœuds de coefficient ayant au moins un voisin significatif pour le seuil précédent et finalement un pour encoder les nœuds de pixel isolées.

Seulement deux contextes pour la passe de raffinement sont utilisés, un pour les nœuds de pixel ayant des voisins non significatifs pour le seuil précédent, un pour les nœuds de coefficient ayant

au moins un voisin significatif pour le seuil précédent.

1.5.3.2 QTL 3D

QTL 3D est un algorithme de codage [103] basé également sur un partionnement en blocs de la TO3D qui est une extension de la méthode QTL[12] 2D . De façon similaire au Cube Splitting (CS), QTL 3D construit des arbres en octons correspondant à chaque carte de signification. La même règle de découpe que dans CS est utilisée de façon récursive sur des cubes de la TO3D. On sélectionne encore une fois des ensembles de bits dans la carte de signification, c'est-à-dire des ensembles de coefficients significatifs de la TO3D. L'encodage de la carte de signification (c'est-à-dire les positions des coefficients significatifs) est équivalent à l'encodage des arbres auxquels ils appartiennent.

La première différence notoire avec CS est que le processus de partitionnement est limité de telle façon, qu'une fois le volume du nœud plus petit qu'un volume seuil prédéfini, le processus de découpe est stoppé et le codage entropique des coefficients à l'intérieur de la feuille significative est activé.

La seconde différence entre les deux méthodes est l'ajout d'une passe supplémentaire. A part les deux étapes de codage classique, c'est-à-dire la passe de signification et la passe de raffinement, une nouvelle étape de codage appelée la passe non significative est introduite. De façon basique, pendant la passe de signification d'un plan de bits donné, les cordonnées des coefficients trouvés comme non significatifs dans un noeud significatif sont ajoutées dans une liste de coefficients non significatifs (LNC). Pendant, les étapes suivantes de codage la signification des coordonnés enregistrées dans la LNC est codée en premier. Ce choix est motivé par les remarques suivantes :

- Les coefficients enregistrés dans la LNC sont localisés dans le voisinage des coefficients qui ont déjà été classés comme significatifs à l'étape de codage courante ou précédente.
- Il y a une forte probabilité pour ces coefficients de devenir significatifs à la prochaine étape de codage du fait de la propriété de d'agglutinement sur laquelle les codeurs par découpe sont basés.

La troisième différence essentielle par rapport au codeur CS est que la phase de contexte et le codage entropique basé sur contexte des symboles générés pendant les 3 passes de codage sont plus élaborés. On va utiliser codage arithmétique basé sur contexte plus compliqué pour améliorer la performance de codage car les modèles simples nécessitant peu de mémoire ne sont parfois pas assez efficaces.

Quatre ensembles différents de modèles sont utilisés pour encoder les symboles générés par cet algorithme de codage et l'encodeur sélectionne automatiquement l'ensemble approprié à chaque phase de codage.

1.5.3.3 CS EBCOT

L'algorithme CS-EBCOT [103] est une association des algorithmes CS [101] et d'une version 3D de EBCOT [106] qui est l'algorithme de codage de JPEG2000.

Dans un premier temps, on découpe les coefficients d'ondelettes comme le fait EBCOT en des cubes séparés de mêmes tailles : les "code-blocs" B. Généralement, les "code-blocs" sont des éléments de taille $64 \times 64 \times 64$. On peut également admettre d'autres tailles (y compris une différente pour chaque dimension). Cela reste conditionné cependant à la taille de l'image et aux caractéristiques de l'application. Comme EBCOT, CS-EBCOT comprend 2 parties principales :

- Niveau 1 : association imbriquée du CS et d'un codage par plan de bits en plusieurs passes grâce à un codeur arithmétique basé sur contexte identique à EBCOT.
- Niveau 2 : formation de paquets et de couches.

[12] *QuadTree-Limited.*

La passe de partitionnement en cubes provient donc de l'algorithme CS. On applique le CS sur des "code-blocs" individuels pour isoler les sous-cubes contenant potentiellement des coefficients d'ondelettes significatifs. A la manière de QTL 3D, on ne va pas jusqu'aux simples coefficients d'un "code-bloc". La taille minimale qui peut être tolérée pour les petites entités est de $4 \times 4 \times 4$. Ainsi, ces plus petits sous-cubes deviennent les nœuds feuilles.

Pendant la première passe, classiquement, on teste toujours à l'aide du même opérateur de signification σ la signification du "code-bloc" B pour le bit le plus significatif, c'est à dire pour le plus grand plan de bits. Dans le cas où $\sigma_{T_n}(B) = 1$, "le code-bloc" est découpé jusqu'à ce que les nœuds feuilles soient isolés. Au moment où la totalité des nœuds feuilles significatifs sont isolés, on active la partie codage par plan de bits pour le plan de bit courant et seulement pour les nœuds feuilles significatifs. Une fois le plan de bits courant codé, il faut passer au plan de bits suivant et réappliquer CS. On continue cette procédure jusqu'à ce que l'on ait codé totalement le "code-bloc"

Intéressons-nous à présent au codage des plans de bits une fois que la découpe du code-bloc est faite. Le codeur de plans de bits ne code que les nœuds feuilles identifiés comme significatifs pendant la passe CS. Trois passes sont définies par plan de bits comme dans le cas 2D. Ces 3 passes illustrées dans la figure 1.17 sont les suivantes :

- La passe de signification : C'est donc la première passe à chaque nouveau plan de bits. Un symbole est codé si il est non significatif, mais qu'au moins un de ces 26 voisins connectés (8 autour de lui en 2D et 9 dans chaque coupe devant et derrière) est significatif. Les symboles significatifs ne sont pas affectés par cette passe. On peut rappeler qu'un symbole significatif a la valeur 1.
- La passe de raffinement : Dans le plan de bits donné, tous les symboles qui sont devenus significatifs dans les plans de bits précédents sont codés (on exclut les symboles qui viennent juste de devenir significatif)
- La passe de normalisation ou de nettoyage : Dans le plan de bits donné, tous les symboles restant sont codés.

De plus, ces 3 passes vont appeler différentes opérations de codage (les primitives) : Le codage de zéro, le codage de signe, le raffinement de magnitude et le codage run-length [52]. Un codage run-length utilise un code qui est appliqué pour l'exploitation des plages de zéros. L'idée du run-length est de ne pas coder indépendamment tous les zéros à la suite mais plutôt le nombre de zéros par plage (appelée aussi runs), ainsi que les valeurs non nulles (appelées levels).

Ces primitives permettent la sélection de modèles de contexte appropriés pour le codage arithmétique à venir et pour les étapes de codage run-length.

En résumé pour chaque plan de bits, successivement la passe de signification, la passe de raffinement de magnitude, la passe de cube-splitting et la passe de normalisation sont appelées, à l'exception du premier plan de bits où les deux premières passes sont supprimées. La figure 1.17 illustre bien ces quatre passes. Cette figure montre en gris foncé, les 26 voisins du coefficient en noir qui est significatif. Ces 26 coefficients sont codés pendant la passe de signification. Puis, à la passe de raffinement seul le coefficient en noir est codé du fait qu'il était déjà significatif. Par ailleurs, on remarque (en gris clair) tous les autres coefficients du volume significatif qui sont codés pendant la phase de nettoyage. Enfin, apparaissent en blanc sur la figure tous les autres coefficients qui n'ont pas été marqué comme significatifs dans les arbres produits par CS et qui ne sont pas codés.

La première étape a généré par le codage arithmétique un ensemble de flux binaires indépendants (un par "code-blocs"), chacun des flux étant imbriqués du fait de la notion de plan de bits. La deuxième étape, qui est identique au celle du standard JPEG2000, va être chargée de multiplexer ces flux en un flux global de codage et de donner l'ordre dans lequel ce flux est structuré. La connaissance de cet ordonnancement est en effet essentielle pour la fonctionnalité de transmission progressive. Cette étape va produire des couches et des paquets qui seront transmis. Nous ne détaillerons pas la construction de ces éléments. Le lecteur intéressé pourra consulter [73] et [107] pour des détails complémentaires sur l'organisation en couches et paquets.

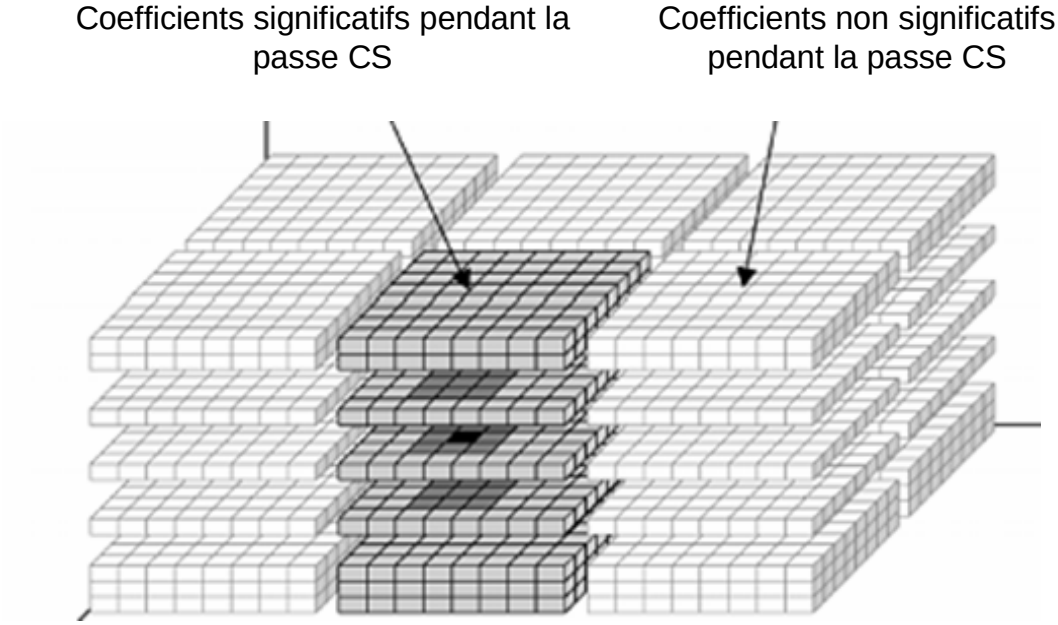

FIG. 1.17: Représentation d'un plan de bits. Pour chaque plan de bits, la passe de signification, la passe de raffinement, la passe de partitionnement en cubes et la passe de normalisation sont appelées successivement sauf pour le premier plan de bits où les deux première passes sont enlevées.

On peut également noter que cette seconde étape de CS EBCOT comprend une procédure d'optimisation de la distorsion de post-compression. On entend par post-compression le fait que l'allocation de débits qui cherche à minimiser la distorsion pour chaque "code-bloc" est calculée après codage de l'image toute entière (en conservant seulement certains plan de bits).

1.5.3.4 ESCOT 3D

ESCOT 3D [117] est également une extension 3D de EBCOT [106], [108] ressemblant ainsi fortement à CS EBCOT sauf sur la partie partitionnement en cubes significatifs du CS. Il garde le concept de plans de bits fractionnaires codés par les 3 passes présentées précédemment : passe de signification, passe de raffinement et passe de normalisation. La principale contribution de ESCOT 3D dans sa version pour l'imagerie médicale volumique[118] est liée à la formation et la modélisation des contextes 3D pour le codage arithmétique des plans de bits. Nous rappelons que réaliser un modèle de contexte 3D consiste à estimer la probabilité conditionnelle de $P(w|C)$ où w est le coefficient d'ondelettes compressé et C un contexte causal approprié.

A l'intérieur d'une coupe donnée, ESCOT 3D utilise le même contexte que celui utilisé par la version d'EBCOT. Le contexte comprend les huit voisins connectés du coefficient d'ondelettes courant. De plus, comme dans CS EBCOT, ESCOT 3D étend les coefficients à prendre en compte pour le modèle de contexte C aux coupes voisines (précédente et suivante) pour exploiter les dépendances dans la troisième dimension (axe z)

Mais pour prévenir le problème de dilution de contexte déjà évoqué pour CB EZW 3D, ESCOT 3D réduit le nombre de coefficients à prendre en compte dans le contexte 3D C. A la place d'ajouter à C des coefficients individuels des coupes situées avant et après la coupe courante, l'algorithme les fusionne dans un modèle simple qui capture la corrélation entre le coefficient w à coder et ses voisins des coupes précédentes et suivantes. Ceci est réalisé via un estimateur linéaire Δ de l'amplitude de w. Notons que c'est une différence avec CS-EBCOT où les 18 voisins liés à l'axe z sont pris en compte pour le contexte (figure 1.17).

On note $w(i_0, j_0, k_0)$ le coefficient d'ondelettes localisé en (i_0, j_0, k_0) et on considère l'ensemble S des 18 voisins de $w(i_0, j_0, k_0)$ des coupes situées avant et après la coupe courante :

$$S = \left\{ \; w(i, j, z_0 - 1), w(i, j, z_0 + 1) \mid i_0 - 1 \leq i \leq \; i_0 + 1, \; j_0 - 1 \leq j \leq \; j_0 + 1 \; \right\} \tag{1.48}$$

L'algorithme utilise une régression linéaire pour calculer α_k, $0 \leq k < 18$ qui minimise :

$$E = \left\{ \left\| \sum_{w_k \in S} \alpha_k \left| w_k \right| - w(i_0, j_0, k_0) \right\|_2 \right\} \tag{1.49}$$

L'estimateur des moindres carrés de $w(i_0, j_0, k_0)$, $\Delta = \sum_{w_k \in S} \alpha_k \left| w_k \right|$ réduit la dimension de l'espace de contexte de 18 (pour S) à 1 (pour Δ). Maintenant, $P(w(i_0, j_0, k_0)|\Delta)$ sera estimé plutôt que $P(w(i_0, j_0, k_0)|S)$ dans la modélisation de la dépendance du coefficient avec les coupes précédentes et suivantes.

Un avantage de la régression linéaire dans un espace de contexte est qu'elle permet une formation de contexte optimisée sous le critère de minimum d'entropie conditionnelle. Puisque Δ est une variable aléatoire scalaire, on conçoit un quantificateur de contexte Q pour minimiser l'entropie conditionnelle $H(w|Q(\Delta)) = E\left\{\log \; P(w|Q(\Delta))\right\}$ à travers un traitement de programmation dynamique.

Au final, le quantificateur de contexte Q réduit l'estimateur Δ sur un nombre modeste K d'états de codage inter-coupes. Dans [118], $K = 4$ apparaît comme suffisant. Le contexte inter-coupes proposé $Q(\Delta)$ et le contexte intra-coupes C sont combinés à travers un produit cartésien dans le codage arithmétique adaptatif de w. A savoir, la probabilité conditionnelle $P(w|C, Q(\Delta))$ est estimée à la volée et utilisée pour produire le code arithmétique pour comprimer w.

Notons enfin que deux processus d'optimisation séquentielle sont nécessaires dans le schéma de modélisation de contexte 3D de ESCOT 3D. La premier concerne la détermination des coefficients de la régression linéaire α_k dans (1.49). Le second est lié à la conception du quantificateur pour l'entropie conditionnelle minimale.

Le nouvel algorithme présenté dans le chapitre 2 se différencie de ces approches intra-bandes. Il est réellement intra-bandes car il effectue une véritable allocation de bits pour chaque sous-bande produite par la TO3D sous la contrainte de minimiser la distorsion totale dans le domaine des ondelettes. Au contraire de toutes les autres méthodes présentées qui sont basées sur des codeurs performants par plan de bits, notre approche est davantage centrée sur une quantification optimale que sur le codage.

1.6 Comparaisons des performances

En préambule, il est important de souligner que la comparaison des performances fournies par toutes les méthodes de compression n'est pas une tâche facile. En effet, contrairement à la compression d'images naturelles, il n'existe pas de bases de données dites de référence pour tester les méthodes. Ainsi, les chercheurs utilisent leurs propres images médicales pour évaluer leurs algorithmes. Or un résultat valable pour une modalité et un organe ne l'est pas nécessairement pour une autre modalité ou un autre organe. Cette partie essaie cependant de résumer les principaux résultats en compression sans et avec perte qui ont été publiés à ce jour.

Modalité	Etude / Nom du volume	Age	Nom Fichier	Taille volume	Taille voxels (mm)
Scanner	fracture du pied / CT1	16	CT_skull	$256 \times 256 \times 192$	$0.7 \times 0.7 \times 2$
Scanner	fracture scaphoïde / CT2	20	CT_wrist	$256 \times 256 \times 176$	$0.17 \times 0.17 \times 2$
Scanner	dissection carotide / CT3	41	CT_carotid	$256 \times 256 \times 64$	$0.25 \times 0.25 \times 1$
Scanner	syndrome d'Apert / CT4	2	CT_Aperts	$256 \times 256 \times 96$	$0.35 \times 0.35 \times 2$
IRM	Foie normal / MR1	38	MR_liver_t1	$256 \times 256 \times 48$	$1.45 \times 1.45 \times 5$
IRM	Foie normal / MR2	38	MR_liver_t2el	$256 \times 256 \times 48$	$1.37 \times 1.37 \times 5$
IRM	exopthalmoses gauches / MR3	42	MR_sag_head	$256 \times 256 \times 48$	$0.98 \times 0.98 \times 3$
IRM	Maladie de coeur / MR4	1	MR_ped_chest	$256 \times 256 \times 64$	$0.78 \times 0.78 \times 5$

TAB. 1.1: Description des 8 images de l'étude E1.

1.6.1 Compression sans perte

De nombreuses études [20], [34], [59], [83] ont testé les méthodes de référence 2D (CALIC) et les standards de compression actuels (JPEG sans perte, JPEG - LS, JPEG 2000, PNG ...)

Kivijärvi *et al* dans [59] examinent la compression sans perte sur 3147 images de différentes modalités (scanner, IRM, échographie, PET, SPECT) . CALIC [115], [116] qui est un codeur prédictif basé sur contexte donne les meilleurs performances en un temps raisonnable (TC = 2,98 : 1 en moyenne) alors que JPEG-LS [114] est presque aussi efficace (TC = 2,81 : 1 en moyenne) et beaucoup plus rapide (4 fois plus rapide). PNG et JPEG sans perte apparaissent en retrait.

Clunie dans [20] teste les mêmes méthodes en évaluant en plus le standard JPEG2000 [67]. Sur 3679 images de différentes modalités et parties anatomiques, il arrive à la conclusion que CALIC est légèrement supérieur (TC = 3,91 : 1 en moyenne) à JPEG LS et JPEG2000 (TC = 3,81 en moyenne pour les 2). JPEG-LS est cependant plus simple à implanter, consomme moins de mémoire et est plus rapide que JPEG 2000 et CALIC. JPEG 2000 possède lui le plus de fonctionnalités pour l'imagerie médicale. Les autres algorithmes testés (LZW, Huffman adaptatif, JPEG sans perte, PNG, Unix compress) présentent de mauvaises performances.

En résumé, l'ensemble des études 2D confirme que JPEG LS, JPEG 2000 et CALIC constituent l'état de l'art pour les algorithmes 2D de compression sans perte d'images médicales. Cependant, remarquons que ces taux de compression restent limités. L'utilisation de méthodes 3D va ainsi permettre d'augmenter légèrement les performances de compression sans perte. L'ensemble de ces techniques correspond aux méthodes présentées dans le paragraphe précédent. Elles sont toutes basées sur une transformée en ondelettes 3D.

Pour faciliter les comparaisons, quatre travaux importants [10], [19], [58] et [118] ont utilisé une base de données commune pour évaluer les performances de compression. Cette base ne possède cependant que 8 images 3D (4 scanners et 4 IRM) codées sur 8 bits (256 niveaux de gris)[13]. Le tableau 1.1 résume les informations de cette base d'images. La figure 1.18 présente la première coupe des 8 volumes de la base. Nous appellons cette étude E1. Le tableau 1.2 résume l'ensemble des résultats obtenus dans ces publications. Il présente pour chaque méthode le débit binaire (en bits par voxel) obtenu pour comprimer chacune des 8 piles d'images. Dans ces quatre publications, les auteurs font varier plusieurs paramètres pour tester leur algorithme de compression. Ils font à chaque fois évoluer les filtres utilisés, le nombre de coupes à rassembler pour les compresser (groupe de coupes) et le nombre de décompositions de la TO3D suivant les dimensions (sauf dans [58]). De plus dans [118], le type de normalisation varie. Dans un souci de concision, nous retiendrons seulement le meilleur résultat publié dans chaque cas. Il apparaît que 3D ESCOT avec une TO3D par paquets d'ondelettes utilisant une normalisation par décalage de bits est la plus performante pour l'ensemble des 8 images 3D. Il est suivi de près par AT - SPIHT 3D (version avec des arbres

[13] Ces images sont téléchargeables à l'adresse *www.cipr.rpi.edu/resource/sequences/sequence01.html*

	CT1	CT2	CT3	CT4	MR1	MR2	MR3	MR4
Unix Compress	4,13	2,72	2,78	1,73	5,30	3,93	3,59	4,33
JPEG2000 [107]	3,08	1,79	1,98	1,28	3,40	2,57	2,81	3,10
SPIHT 2D [98]	2,69	1,84	1,98	1,23	3,37	2,58	2,75	2,90
JPEG LS [114]	2,84	1,65	1,73	1,06	3,16	2,37	2,56	2,93
CALIC	2,72	1,69	1,65	1,04	3,05	2,24	2,58	2,81
EZW 3D [10]	2,22	1,28	1,50	1,00	2,37	1,80	2,39	2,04
CB EZW 3D [10]	2,01	1,14	1,39	0,89	2,20	1,65	2,28	1,87
SPIHT 3D [58]	1,95	1,14	1,46	0,93	2,24	1,67	2,07	1,72
AT - SPIHT 3D [19]	1,92	1,11	1,48	0,91	2,20	1,64	**1,92**	1,65
ESCOT 3D [118]	**1,83**	**1,05**	**1,34**	**0,86**	**2,08**	**1,51**	1,94	**1,62**

Tab. 1.2: Comparaison des performances en bits/voxel des différentes techniques de compression sans perte de l'étude E1.

non équilibrés de SPIHT 3D) sauf pour les scanners de la carotide et du syndrome d'Apert où CB EZW 3D est second. Dans tous les cas, les algorithmes 3D surpassent toutes les méthodes 1D (Unix compress) et 2D reconnues en imagerie médicale. Par exemple, la meilleure méthode de ce regroupement d'études ESCOT 3D permet de gagner de 30 à 35 % sur la taille du fichier par rapport à la référence 2D CALIC.

Par ailleurs, une autre étude (nommée ici E2) [102], [103] a testé plusieurs algorithmes sur différentes modalités. Les types d'images médicales testés sont un PET ($128 \times 128 \times 39 \times 12$ bits) , 2 scanners - CT1 ($512 \times 512 \times 100 \times 12$ bits), CT2 ($512 \times 512 \times 44 \times 12$ bits), une échographie ($256 \times 256 \times 256 \times 8$ bits) et une IRM ($256 \times 256 \times 200 \times 12$ bits). Cette étude compare CS, CS-EBCOT, QTL 3D, SPIHT 3D, JPEG 2000 avec une TO3D et JPEG2000. Pour l'échographie et le PET, le codeur 3D QTL délivre les meilleures performances alors que CS-EBCOT est le meilleur pour les 3 autres examens. Ces deux codeurs sont très proches car la différence moyenne entre les deux (en terme de taux de compression) n'est que de 0,1%. SPIHT 3D et JPEG 2000 avec une TO3D présentent une différence de 3,65% et 3,67 % en moyenne par rapport à la meilleure méthode pour l'image testée. Seul JPEG2000 (avec une TO2D) est testé comme méthode 2D et obtient les plus mauvais résultats à l'exception du CT2 car la résolution en z est limitée.

Pour résumer, l'ensemble de ces publications permet de faire les observations suivantes pour le codage sans perte :

- L'utilisation d'une **transformée en ondelettes 3D** en présence d'une pile d'images importante augmente sensiblement les performances de codage. Les méthodes 3D sont toujours meilleures que les méthodes 2D.
- L'approche intra-bande 3D présente de meilleurs résultats que l'approche inter-bande 3D pour le codage sans perte, comme en témoignent ESCOT 3D dans la première étude et QTL 3D et CS - EBCOT dans la seconde étude présente les meilleurs résultats. Ceci est logique car l'efficacité des approches inter-bandes basées sur l'hypothèse des arbres de zéros décroît au fur et à mesure des plans de bits. Or la compression sans perte nécessite la transmission de tous les plans de bits.

Néanmoins, l'utilisation des méthodes 3D ne permet pas d'atteindre de très forts taux de compression lorsqu'il s'agit de compression sans perte. La compression avec pertes maîtrisées peut donc être une réponse à cette limitation.

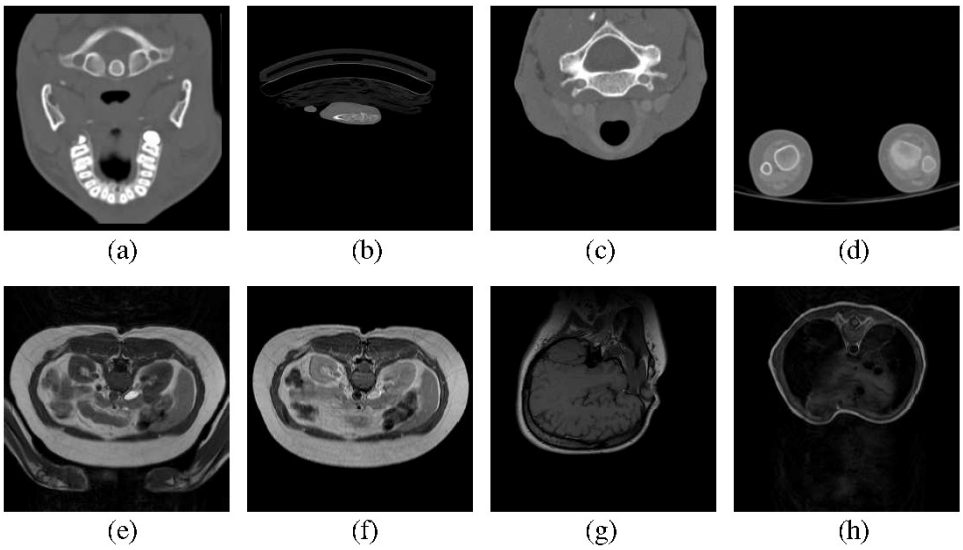

FIG. 1.18: Images médicales volumique de la base E1. Première coupe de chaque pile. (a) *CT_skull* (b) *CT_wrist* (c) *CT_carotid* (d) *CT_aperts* (e) *MR_liver_t1* (f) *MR_liver_t2_el* (g) *MR_sag_head* (f) *MR_ped_chest*.

1.6.2 Compression avec perte

Comparer les algorithmes de compression avec pertes en imagerie médicale est un problème très complexe. Dans l'introduction, nous avons évoqué les difficultés pour évaluer les pertes, la qualité de l'image reconstruite, la précision du diagnostic. Dans le même temps, nous avons décrit brièvement un certain nombre d'outils statistiques et subjectifs permettant d'évaluer les performances des méthodes. Malheureusement, la quasi totalité des publications actuelles de référence en compression d'images médicales n'utilisent pas ces techniques car elles sont souvent très coûteuses en temps et en moyens humains. De plus, il n'y a à notre connaissance presque aucun travail récent où l'évaluation subjective de la qualité des images compressées a été réalisée par des professionnels de la radiologie. Le PSNR demeure le critère essentiel d'évaluation alors que cette grandeur mathématique constitue une mesure de qualité souvent inadéquate pour les images comprimées avec perte, spécialement d'ailleurs pour les applications médicales. Ainsi, les études comparatives que nous résumons ici présentent des résultats s'appuyant sur le PSNR et l'évaluation subjective des auteurs.

Les images de la base de l'étude E1 ont également été comprimées avec perte dans trois études précitées [10], [19], [118]. Il est par contre plus difficile à partir de ces articles de synthétiser tous les résultats. Cependant, dans [118], SPIHT 3D apparaît meilleur que SPIHT 2D respectivement de 7,65 dB et 6,21 dB à 0,1 bit/voxel et 0,5 bit/voxel. ESCOT 3D permet encore un gain de 0,69 dB et 0,93 dB par rapport à SPIHT 3D. Si on compare ces résultats avec ceux fournis dans [10], le codage avec perte résultant de ESCOT 3D est 3,0 et 4,0 dB meilleur que CB EZW 3D à 0,1 bit/voxel et 0,5 bit/voxel respectivement. Les auteurs ne donnent pas d'appréciation sur la qualité visuelle de la reconstruction mais comparent la complexité. Il apparaît que ESCOT 3D a une complexité 50 % supérieur à celle de SPIHT 3D.

Dans l'étude E2 [102], [103], les résultats sont donnés pour deux types de filtre pour la TO3D : le filtre 5.3 sans perte et le filtre 9.7 avec perte. Pour le filtre 5.3, QTL 3D délivre les résultats les plus intéressants en terme de PSNR. Mais pour le filtre reconnu le meilleur pour la compression avec perte, à savoir le 9.7 de Daubechies, SPIHT 3D délivre les meilleures performances en terme de PSNR suivi de près par CS et CS-EBCOT. L'étude E2 conclue également que l'évaluation subjective de la qualité visuelle de reconstruction démontre que les techniques 3D basées sur les on-

delettes délivrent des résultats sensiblement équivalents. Notons simplement que CS et CS-EBCOT préservent un tout petit peu mieux les structures spatiales basse-fréquences.

Pour résumer l'ensemble des résultats issus de la **compression avec pertes**, ESCOT 3D et SPIHT 3D apparaissent comme des références en terme de qualité numérique et visuelle suivis par CS et CS-EBCOT.

1.7 Conclusion

La gestion des données informatiques dans les services hospitaliers est devenue un enjeu majeur pour la mise en place de PACS performants. Le volume de ces données, déjà impressionnant, continuera de croître dans les années à venir du fait du développement de systèmes d'imagerie médicale de plus en plus performants, permettant une investigation de plus en plus fine des organes humains. La compression et particulièrement celle avec pertes maîtrisées peut apporter une réponse à ce problème et permettre un stockage et une transmission plus efficace au sein ou au-delà du PACS. En témoigne l'intégration du nouveau standard JPEG2000 dans le format de gestion des images médicales DICOM et surtout le développement actuel de la partie 10 de JPEG 2000 en cours de construction : JPEG2000 3D[14]. Cette partie reprend de nombreuses notions présentées dans ce chapitre. Elle utilise la transformée reconnue comme la référence pour les images médicales 3D : la transformée en ondelettes 3D. JPEG2000 3D doit également reprendre les principales caractéristiques des meilleurs codeurs 3D comme CS-EBCOT et ESCOT 3D que nous avons décrits précédemment. Parallèlement, la recherche dans le domaine continue de proposer de nouveaux codeurs qui tentent de répondre à la spécificité de l'imagerie médicale. C'est le cas de notre méthode QVAZM 3D qui sera présentée dans le chapitre 2 présentant sur plusieurs types d'images des résultats supérieurs à bas débits à ceux obtenus avec SPIHT 3D, l'une des méthodes de référence dans ce domaine. Dans l'état de l'art que nous avons cherché à faire ici, il semble maintenant reconnu qu'une compression performante d'images médicales 3D (piles d'images 2D) passe par une transformée en ondelettes 3D, suivie d'un codeur approprié (inter-bandes ou intra-bandes). Cette structure permet en effet d'exploiter de manière efficace à la fois les corrélations existant au sein d'une coupe, mais également entre les coupes. Si les techniques dont nous nous sommes fait l'écho dans ce chapitre offrent des performances intéressantes en terme de taux de compression, elles souffrent encore de quelques obstacles, parmi lesquels en premier lieu, l'évaluation de la qualité des images comprimées, en particulier lorsque l'enjeu concerne le diagnostic. Le chapitre 3 traitera en partie de l'évaluation de la qualité en étudiant certains effets de la compression sur un outil d'aide à la décision pour les images médicales. De plus, l'absence de base d'images médicales de référence qui pourrait offrir un standard doré et permettre une comparaison rigoureuse des méthodes ne facilite pas non plus l'évaluation de performances. Quoiqu'il en soit, même s'il existe de nombreux outils statistiques d'analyse de performances, l'appréciation de la qualité des images reste du ressort des médecins, ce qui rend les tests à grande échelle difficiles à réaliser.

Pour conclure, nous insistons cependant sur le fait que la théorie de l'information imposant des limites à la compression sans perte, celle-ci n'offre des taux que très modestes (par rapport aux enjeux), et ce malgré des codeurs de plus en plus sophistiqués. La compression avec pertes - à condition bien entendu que celles-ci soient maîtrisées et compatibles avec les applications médicales - apparaît comme la réponse la plus appropriée au problème d'archivage et de transmission des données médicales, dont les volumes augmentent de façon exponentielle.

[14] http ://www.jpeg.org/jpeg2000/j2kpart10.html

Chapitre 2

Compression par Quantification Vectorielle Algébrique avec Zone Morte 3D pour l'imagerie radiologique

2.1 Introduction

Le chapitre précédent a présenté les enjeux liés à la compression des images médicales volumiques, les principales méthodes 3D existantes et la potentialité des deux types de compression (sans perte, avec pertes). Il a en outre montré que les meilleures techniques 3D de compression sans perte permettaient d'atteindre des taux de compression proches de 8 : 1 au maximum pour le panel d'images de la base E1. Cependant, ces taux ne sont pas compatibles avec de nombreuses applications (télémédecine, recherche rapide, système de grilles pour l'imagerie médicale [42] et navigation de données médicales volumiques). C'est pourquoi pour certaines applications, la compression avec pertes parait inévitable et doit être explorée avec attention.

Dans ce chapitre, nous proposons une nouvelle méthode de compression avec pertes dédiée aux images médicales volumiques. Notre chaîne de compression s'appuie en premier lieu sur la transformée reconnue comme une référence pour les piles d'images médicales : la Transformée en Ondelettes 3D (TO3D). Notre approche se situe dans la famille des méthodes intra-bandes où les sous-bandes sont codées de manière indépendante par un algorithme d'allocation de débits. La seconde étape correspondant à la quantification utilise une Quantification Vectorielle Algébrique (QVA). La principale contribution de ce travail est la conception d'une zone morte multidimensionnelle pendant l'étape de quantification pour prendre en compte les corrélations entre les voxels voisins. En plus de la TO3D, la notion de 3D dans notre chaîne de compression est présente au moment de la découpe des vecteurs. En effet, l'orientation des vecteurs peut être choisie dans les trois dimensions afin d'avoir une distribution des énergies des vecteurs favorables à la Quantification Vectorielle Algébrique avec Zone Morte (QVAZM). Dans la suite du manuscrit, notre algorithme sera nommé QVAZM 3D. La figure 2.1 résume les étapes de la chaîne de compression proposée.

Par ailleurs, cette zone morte devient un paramètre supplémentaire à régler dans l'allocation de débits. Nous présenterons une méthode de recherche rapide de la zone morte qui s'insérera dans la procédure d'allocation de débits. L'allocation correspondant à un problème classique de minimisation sous contraintes d'égalité s'appuiera sur une méthode d'optimisation numérique de type lagrangienne dans un premier temps[1] qui peut se révéler très couteuse pour des gros volumes de données médicales.

[1] Le Chapitre 4 présentera des méthodes d'optimisation beaucoup plus rapides pour la QVAZM.

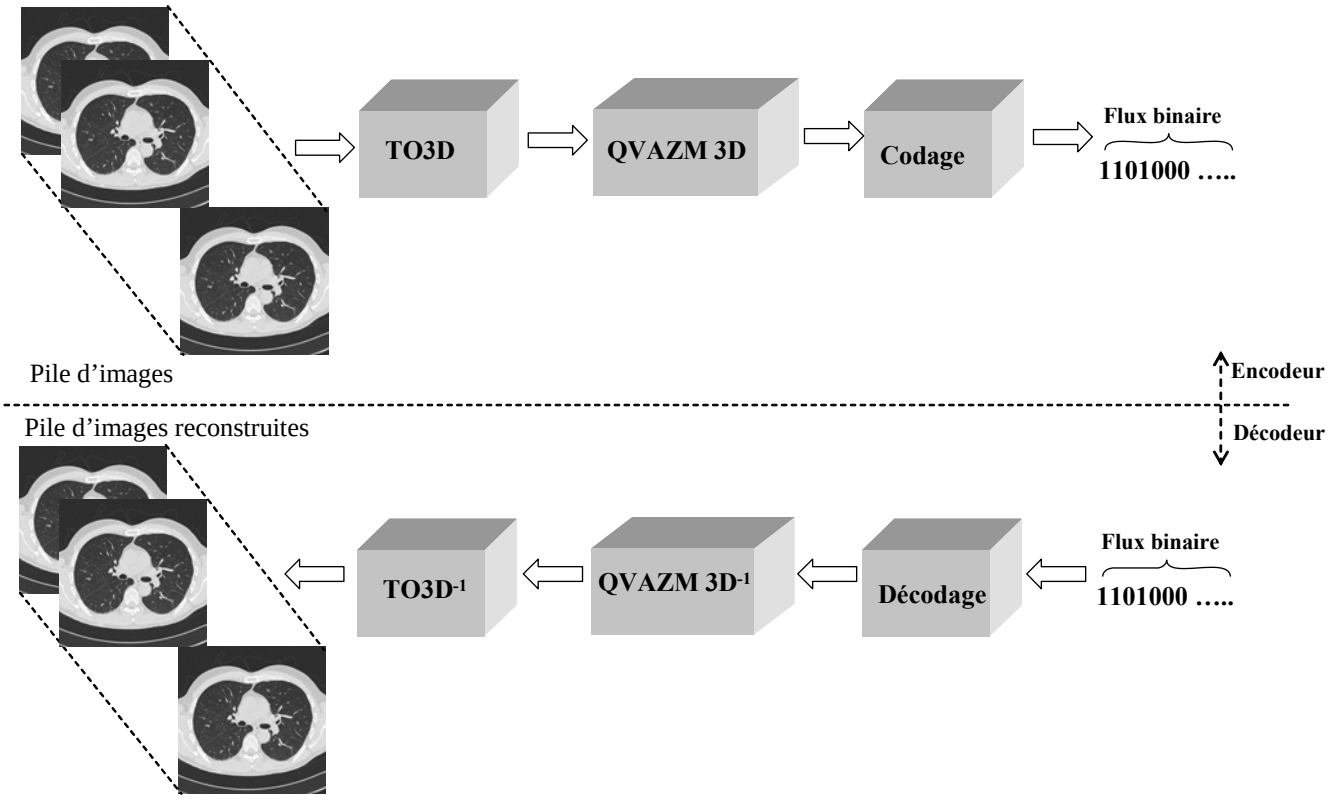

Fig. 2.1: Schéma de compression/décompression de la QVAZM 3D pour les images médicales volumiques 3D.

Notre algorithme a été évalué sur les mêmes scanners et IRM que ceux de la base E1 décrite au chapitre 1. Pour ces données, à fort taux de compression pour l'imagerie médicale mais à qualité visuelle encore acceptable, notre méthode est supérieure aux meilleures méthodes existantes en termes de compromis débit/distorsion et de qualité visuelle. Des simulations supplémentaires de notre méthode ont été effectuées sur des scanners ORL. Elles ont conduit un radiologue et un radiothérapeute à comparer notre méthode à l'une des références en compression d'images volumiques SPIHT 3D.

Ce chapitre rappellera les principales caractéristiques de la Quantification Vectorielle Algébrique (QVA) puis introduira le concept de dépendances spatiales pour les vecteurs des sous-bandes de la TO3D. La Zone Morte vectorielle (ZM) pour l'imagerie médicale sera présentée dans le paragraphe 4. Le paragraphe 5 expliquera comment régler cette zone morte et comment résoudre le problème de l'allocation de débits. Le dernier paragraphe exposera les résultats expérimentaux.

2.2 La quantification vectorielle algébrique

Rappelons dans un premier temps la définition de la quantification, en insistant sur le fait que l'espace dans lequel elle s'applique est multidimensionnel. Soit $(X_1, ..., X_N)$ un signal échantillonné, avec $X_i \in \mathbb{R}\ \forall i$. En considérant la suite de vecteurs $\left(X^k\right)_{k=1,...,\frac{N}{n}}$ telle que $X^k = \left(X_{n(k-1)+1}, ..., X_{n(k-1)+n}\right)$, on obtient une représentation vectorielle de la source. La quantification du vecteur X de dimension n constitue le passage du continu au discret : il s'agit de projeter X dans un sous-ensemble de $\mathbb{R}^n$

dénombrable et borné, donc fini. Ce passage s'effectue grâce à une application, notée Q :

$$\begin{aligned} Q : \mathbb{R}^n &\to C \\ X &\mapsto Y \end{aligned} \tag{2.1}$$

avec $n \in \mathbb{N}$.

$C \subset \mathbb{R}^n$ est appelé dictionnaire, c'est l'ensemble dénombrable contenant les valeurs que peut prendre le signal quantifié. Lorsque $n = 1$, le quantificateur est dit scalaire (QS), le dictionnaire est un ensemble discret de valeurs réelles et la quantification s'effectue échantillon par échantillon. En revanche lorsque $n > 1$, le quantificateur est dit vectoriel (QV) : le dictionnaire est composé de vecteurs. La QV consiste donc à traiter conjointement un ensemble (généralement connexe) de n échantillons appelé bloc ou vecteur. Une fois la QV effectuée, on obtient un ensemble de $\frac{N}{n}$ vecteurs (on suppose que N est un multiple de n) constituant le signal quantifié.

L'application Q est généralement une projection orthogonale afin de minimiser une distance d entre le vecteur de la source et le vecteur quantifié. L'ensemble V_j des vecteurs de distance minimale du $j^{ème}$ représentant du dictionnaire est appelé cellule de Voronoï, c'est la généralisation de la cellule de quantification dans le cas scalaire :

$$V_j = \left\{ X \in \mathbb{R}^n / d\left(X, Y^j\right) \leq d\left(X, Y^k\right), \forall Y^k \in C,\ k \neq j \right\} \tag{2.2}$$

Généralement, d est la distance euclidienne (norme L^2) afin de minimiser l'erreur quadratique moyenne (EQM).

La supériorité théorique de la QV sur la QS est un résultat bien connu : elle est à la base de la théorie de la distorsion de Shannon [43] [45] [107]. En revanche, la mise en oeuvre de la QV souffre de la taille importante des dictionnaires employés. De nombreuses méthodes ont été proposées dans la littérature, que l'on peut classer selon le mode de partitionnement en cellules de Voronoï sur lequel repose leurs dictionnaires : les partitionnements irréguliers pour les dictionnaires non structurés et les partitionnements réguliers correspondants aux dictionnaires structurés.

Les premiers sont construits à partir des sources à quantifier : une phase d'apprentissage détermine les "meilleurs représentants" pour un ensemble de signaux donné, le dictionnaire ainsi construit doit alors être transmis au décodeur. La figure 2.2 donne un exemple d'un dictionnaire obtenu par apprentissage. La plus célèbre de ces méthodes est le l'algorithme LBG(du nom de ses auteurs Linde-Buzo-Gray) [61] [43]. Le dictionnaire doit être transmis en l'absence d'a *priori* sur le type d'images. C'est la première limite de l'approche par apprentissage. La seconde limite provient du coût de calcul prohibitif de la phase d'apprentissage du dictionnaire. Cependant, il existe des travaux utilisant ce type de méthodes pour coder des images médicales qui pour une classe donnée (même modalité et même organe) ont tendance à se ressembler. Dans ce cas, le dictionnaire n'a besoin d'être calculé et transmis qu'une seule fois et il est "presque optimal". Cosman et al [24] utilisent des méthodes arborescentes pour réduire la phase de projection dans le dictionnaire. Citons également une approche hybride [71] qui découpe les vecteurs en utilisant la structure arborescente de SPIHT et les codent soit par un représentant du dictionnaire (créé par LBG) quand l'erreur est inférieure à un seuil prédéfini, soit par un codage SPIHT lorsque l'erreur est supérieure au seuil. Notons que ces techniques n'ont pas utilisé la TO3D comme première étape de la chaîne de compression et ne peuvent donc pas être comparées aux meilleures méthodes actuelles.

Le deuxième type de partitionnement conduit à un dictionnaire structuré qui ne nécessite pas de phase d'apprentissage. Il existe un grand nombre de méthodes de quantification vectorielle de ce type [26][43][45] telles que la quantification vectorielle à structure en arbre [26], la quantification vectorielle associée à un codage par treillis [107], la quantification vectorielle hiérarchique, etc. Nous allons ici nous focaliser sur la plus répandue de ces méthodes : la **quantification vectorielle sur réseau régulier de points** ou **quantification vectorielle algébrique**. Dans cette méthode

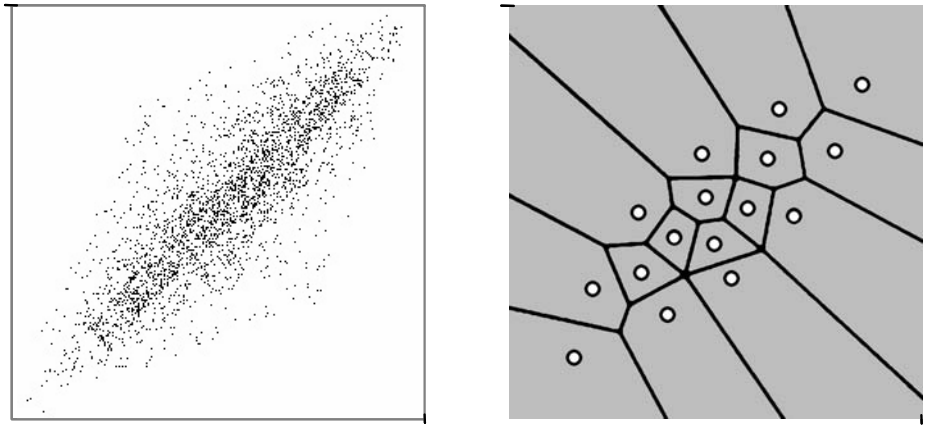

Fig. 2.2: Exemple de dictionnaire non structuré.

le dictionnaire est généré analytiquement, il n'y donc pas de phase de construction et l'étape de quantification a une complexité nettement plus faible que celle des dictionnaires non structurés : elle répond à des opérations mathématiques simples et rapides de type "arrondi". Cattin et al [8] ont été les premiers à notre connaissance à proposer pour les images médicales 3D une méthode basée sur une TO3D suivie par une quantification vectorielle algébrique. C'est également sur cette approche que reposent les travaux présentés dans ce mémoire. La suite du paragraphe concerne les principales étapes de la compression par quantification vectorielle algébrique, à savoir :

- le choix d'un réseau approprié.
- définition du dictionnaire.
- la procédure d'atteinte du débit cible par adaptation de la source.
- le codage des vecteurs produits par la quantification.

2.2.1 Choisir le réseau

Un réseau Λ est un sous-ensemble de $\mathbb{R}^n$ constitué des combinaisons linéaires d'entiers relatifs d'une base de $\mathbb{R}^m$ avec $m \geq n$:

$$\Lambda = \{Y \in \mathbb{R}^m / \exists U \in \mathbb{Z}^n, Y = GU\} \tag{2.3}$$

avec $G \in M^{m \times n}(\mathbb{R})$ la matrice génératrice du réseau.

La recherche de réseaux adaptés à la compression a fait l'objet de nombreux travaux depuis plusieurs décennies. L'effort a d'abord été porté sur la détermination, pour différentes dimensions, des réseaux optimaux *au sens de la théorie haute résolution* [23][43]. Nous allons brièvement décrire les fondements de cette théorie. On suppose que l'on modélise les vecteurs de la source par un vecteur aléatoire X de distribution conjointe f. On introduit le bruit de quantification B défini par :

$$B(X) = X - Q(X) \tag{2.4}$$

avec Q l'opération de quantification sur le réseau Λ.

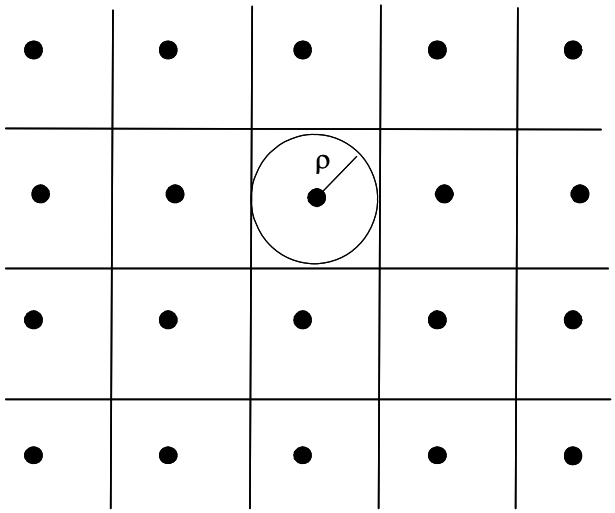

Fig. 2.3: Rayon d'empilement ρ du réseau $\mathbb{Z}^2$.

Soit D l'erreur quadratique moyenne ou distorsion définie par :

$$D = \int_{\mathbb{R}^n} \|B(X)\|_2^2 \, f\,(X)\, dX \tag{2.5}$$

Dans le cadre de la théorie haute résolution, on se place dans l'hypothèse où la distribution du bruit de quantification peut être approximée par une distribution uniforme sur chaque cellule de Voronoï :

$$D \simeq \sum_j \frac{P_j}{|V_j|} \int_{V_j} \|X - Q\,(X)\|_2^2 \, dX \tag{2.6}$$

avec V_j la cellule de Voronoï du $j^{ème}$ vecteur du réseau, $|V_j\,|$ désignant le volume de la cellule V_j et $P_j = P\,(X \in V_j) = \int_{\mathbb{V}_j} f\,(X)\, dX$

Un réseau optimal (au sens de la théorie haute résolution) pour une dimension donnée minimise (2.6). Cette optimalité est liée à la géométrie des réseaux par l'intermédiaire de leur rayon d'empilement ρ qui correspond à la moitié de la distance minimale entre deux vecteurs du réseau c'est-à-dire au plus grand rayon possible de sphères disjointes centrées sur chacun des points du réseau. La figure 2.3 représente les cellules du réseau $\mathbb{Z}^2$ dans lesquelles sont inscrites les sphères de rayon ρ. Plus le rayon d'empilement d'un réseau est petit (on dit que le réseau est dense), plus l'erreur quadratique moyenne (2.6) est faible.

La théorie haute résolution est valide lorsque la taille des cellules de quantification est petite par rapport au taux de variation de la densité de probabilité f. Cependant, dans le cadre spécifique d'une quantification après une TO3D, les coefficients d'ondelettes bénéficient d'une part, d'une statistique extrêmement piquée et offrent d'autre part, la possibilité de quantification assez grossière, particulièrement dans les hauts niveaux de résolution. L'approximation haute résolution de la formule (2.6) n'est donc pas toujours valable, en particulier en compression moyen et bas débit : Gao et *al.* ont, en outre, montré dans [41] que le réseau le plus performant en compression dans le domaine des ondelettes est le moins bon au sens de la théorie haute résolution (plus grand rayon d'empilement), à savoir le réseau $\mathbb{Z}^n$. Pour cette raison, ainsi que pour la faible complexité de l'étape de quantification qui lui est associée, **le réseau Z^n est utilisé dans la suite de ce mémoire.**

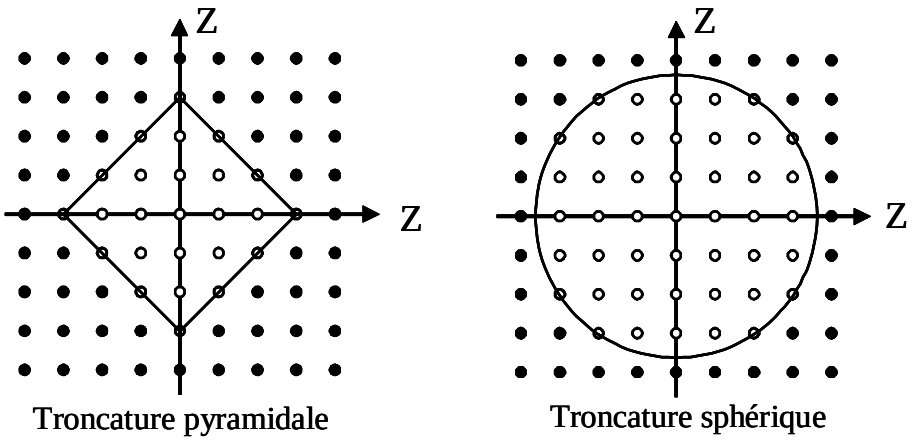

FIG. 2.4: Exemples de troncature du réseau $\mathbb{Z}^n$ pour des sources modélisées par des gaussiennes généralisées de paramètre $\alpha = 1$ et 2.

2.2.2 Définir le dictionnaire

Le dictionnaire C est un sous-ensemble borné du réseau Λ, à présent déterminé. La forme du dictionnaire est fixée en fonction de l'*a priori* sur la statistique des sources à comprimer. Un dictionnaire dont la forme est adaptée à la distribution des vecteurs de la source permet de minimiser le nombre maximal de vecteur requis pour représenter celle-ci, ce qui a une influence directe sur la quantité d'information nécessaire pour décrire la source quantifiée.

Soit r le rayon de l'hypersphère $S_d^n(r)$ définie ci-dessous :

$$S_d^n(r) = \{X \in \mathbb{R}^n / d(0, X) = r\} \tag{2.7}$$

avec d une distance.

Dans l'hypothèse de vecteurs source équirépartis sur les hypersphères S_d^n suivant la distance d (formule 2.7), la forme optimale du dictionnaire est l'hypersphère de rayon de troncature r_T.

La distance d est généralement une norme L^α, $\alpha > 0$. Dans la suite de ce mémoire nous la désignerons indifféremment par les termes : norme ou énergie (ce qui est un abus de langage puisque ce dernier terme est d'ordinaire réservé à la norme L^2).

Dans le cas d'une source i.i.d. dont la distribution suit une loi de type gaussienne généralisée d'hyperparamètre α (représentation classique des coefficients d'ondelettes) , le dictionnaire optimal sera une hypersphère suivant la norme L^α : $S_\alpha^n(r) = \{X \in \mathbb{R}^n / \|X\|_\alpha^\alpha \leq r\}$. La figure 2.4 illustre les cas $\alpha = 1$ et $\alpha = 2$ (distributions respectivement laplacienne et gaussienne). Les dictionnaires correspondant sont alors respectivement de forme pyramidale et sphérique.

Les figures 2.5 et 2.6 présentent les performances de compression par QVA des normes L^1, et L^2 sur les sous-bandes LHL3[2] et LLH3 produites par une TO3D dyadique en utilisant le filtre biorthogonal 9/7 sur le scanner CT1 et sur l'IRM MR2 de la base E1. Nous remarquons que les courbes débit-distorsion sont quasiment identiques. La comparaison des deux normes conduit à des résultats similaires sur toutes les sous-bandes pour l'ensemble des images de la base E1. Ce résultat nous amène à choisir la norme L^1 qui présente la mise en oeuvre la plus simple.

Le dictionnaire est à présent construit, nous allons maintenant détailler les différentes étapes d'une compression par QVA. Celles-ci sont résumées sur la figure 2.7, à savoir : mise à l'échelle de la source, projection sur le réseau (quantification) et codage sans perte des vecteurs composé, d'une part d'un codeur entropique, et d'autre part d'un codage à longueur fixe.

[2]L et H signifie les filtres passe-haut et passe-bas respectivement. La direction du filtrage est dans l'ordre x, y et z. Le numéro à la fin correspond au niveau de décomposition.

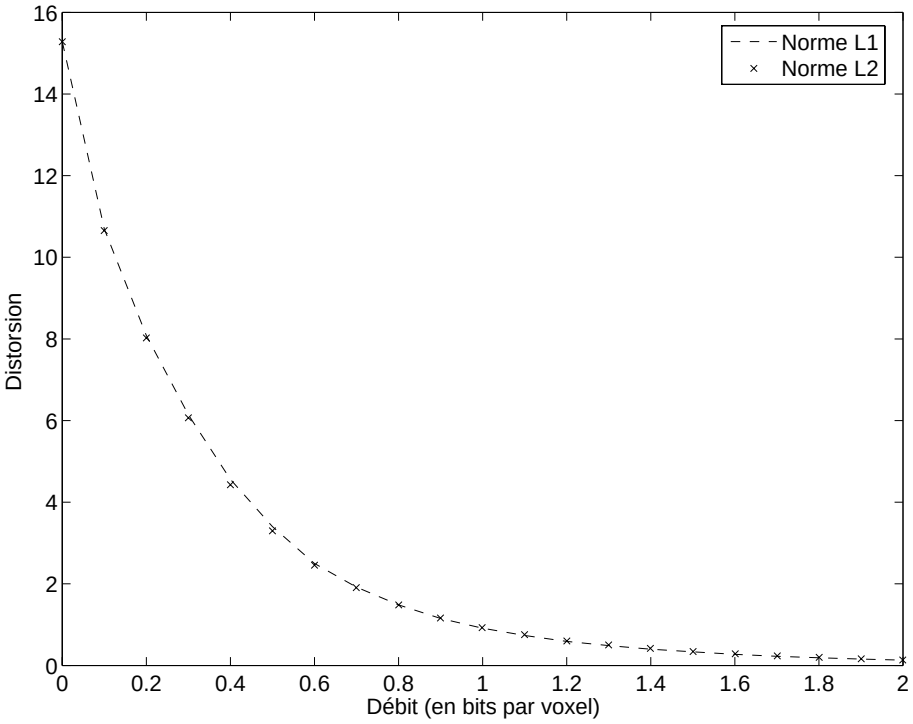

Fig. 2.5: Comparaison des courbes débit-distorsion obtenues par QVA avec la norme L^1et L^2 - sous-bande LHL3 de CT1 - filtre biorthogonal 9.7.

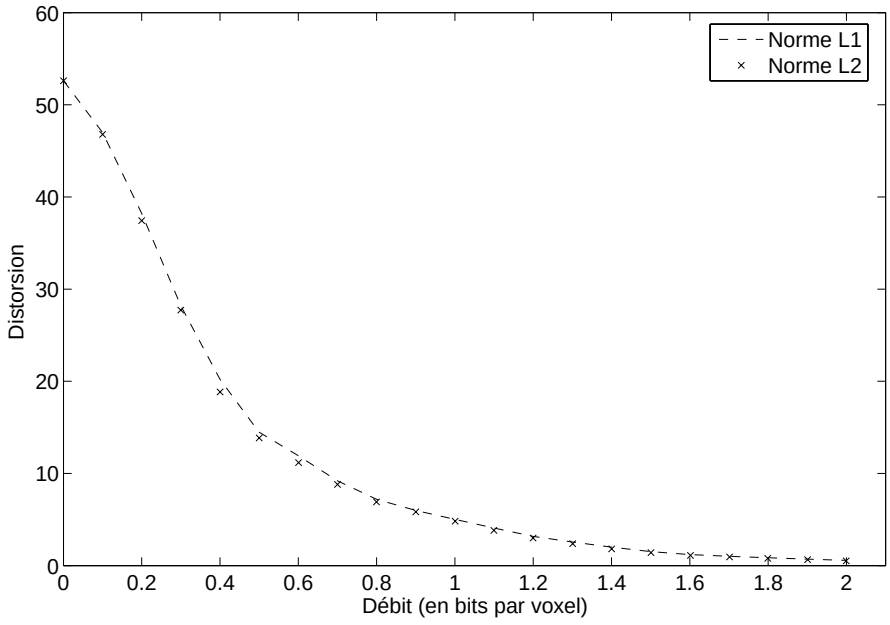

Fig. 2.6: Comparaison des courbes débit-distorsion obtenues par QVA avec la norme L^1et L^2 - sous-bande LLH3 de MR2 - filtre biorthogonal 9.7.

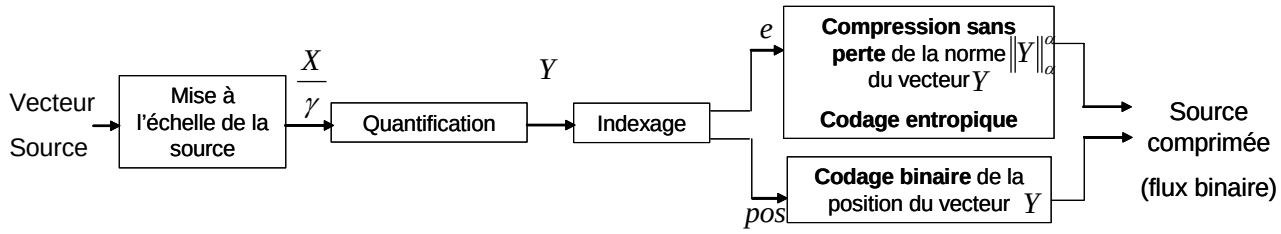

FIG. 2.7: Codage par Quantification Vectorielle Algébrique.

2.2.3 Atteindre le débit cible

Le dictionnaire doit s'adapter à la source en fonction du débit cible : pour cela, la source est projetée à l'intérieur du dictionnaire. Le terme projection est un abus de langage car cette opération est en fait une mise à l'échelle de paramètre γ que nous désignerons dans la suite par facteur de projection :

$$P_\gamma : \mathbb{R}^n \rightarrow \mathbb{R}^n \tag{2.8}$$

$$X \mapsto \bar{X} = \frac{X}{\gamma} \tag{2.9}$$

$\bar{X}$ est appelé le vecteur projeté.

La quantification du vecteur de la source X dans le dictionnaire est la composition entre P_γ et Q la quantification sur le réseau Λ :

$$Y = Q \circ P_\gamma(X) = Q\left(\frac{X}{\gamma}\right) = Q\left(\bar{X}\right) \tag{2.10}$$

Y est un vecteur du réseau, le vecteur quantifié.

La reconstruction des vecteurs après quantification s'effectue de la manière suivante :

$$\tilde{X} = P_\gamma^{-1}\left\{Q\left[P_\gamma(X)\right]\right\} = \gamma Q\left(\frac{X}{\gamma}\right) = \gamma Q\left(\bar{X}\right) \tag{2.11}$$

$\tilde{X}$ est le vecteur reconstruit.

En résumé, pour atteindre le débit cible ou le taux de compression souhaité, il suffit uniquement de régler le paramètre γ. Ce paramètre γ est optimisé pour chaque sous-bandes de la TO3D dans une procédure d'allocation de débits. Ainsi une fois le paramètre régler, ce type de quantification Q est très rapide car il consiste en une simple opération d'arrondi à l'entier le plus proche.

2.2.4 Coder les vecteurs

Le codage décrit ci-dessous repose sur les deux remarques suivantes :

1. L'indexage direct de tous les vecteurs du dictionnaire par l'intermédiaire d'une table de correspondance est irréalisable en raison de son coût calculatoire et de stockage.
2. Dans l'hypothèse d'une source non uniformément distribuée, un schéma de compression doit être capable d'exploiter la redondance induite par la quantification de la source.

Parmi les nombreuses solutions prenant en compte ces deux points [43] [45], les méthodes par codes produit, directement liées à la forme du dictionnaire, semblent les plus adaptées. Ces méthodes sont basées sur la propriété suivante : un vecteur Y d'un réseau peut être caractérisé de manière

unique par sa norme (ou énergie) $e = \|Y\|_\alpha^\alpha$ et sa position *pos* sur l'hypersphère de rayon e (voir définition (2.7)). Le code produit CP est le résultat de la concaténation de ces deux informations :

$$\begin{array}{l} I : CP \to \mathbb{N} \times \mathbb{N} \\ Y \ \mapsto P = (e, pos) \end{array} \tag{2.12}$$

(e, pos) est un mot du code produit, c'est l'index du vecteur Y, e est appelé le préfixe et *pos* le suffixe de l'index.

L'intérêt d'un tel codage réside dans le fait que le code associé à la norme des vecteurs possède une entropie faible qui peut donc être exploitée par un codeur entropique ou par un codeur par plages.

Le calcul du suffixe est quant à lui rendu possible grâce à la donnée du préfixe : au lieu d'indexer l'ensemble des vecteurs du dictionnaire, on ne considère ici que les vecteurs appartenant à l'hypersphère $S_\alpha^n(R)$ de rayon R. La population sur la surface de $S_\alpha^n(R)$ est bien moindre que la population totale du dictionnaire : les références [37], [64] et [75] présentent des algorithmes d'indexage pour la QVA ayant une complexité et un coût de stockage raisonnables, et ce pour différentes formes de dictionnaire et de réseaux.

Le débit R_{TOT} de la source quantifiée est composé de deux termes, le débit de l'énergie DB_E et le débit binaire de la position DB_P :

Débit de l'énergie DB_E

$$DB_E = -\sum_{r=0}^{r_T} P(r) \log_2 (P(r)) \text{ bits/vecteur} \tag{2.13}$$

avec $P(r) = P(e = r) = P(Y \in S_\alpha^n(r))$, Y étant la v.a. représentant la source quantifiée et α la norme associée au mode d'indexage.

Débit binaire de la position DB_P

$$DB_P = \sum_{r=0}^{r_T} P(r) \lceil \log_2 (N(r)) \rceil \text{ bits/vecteur} \tag{2.14}$$

avec $N(r)$ la population de vecteurs de l'hypersphère $S_\alpha^n(r)$.

Le débit total R_{TOT} est donné par la somme de ces deux termes :

Débit total de la source R_{TOT}

$$R_{TOT} = DB_E + DB_P = -\sum_{r=0}^{r_T} P(r) [\log_2 (P(r)) - \lceil \log_2 (N(r)) \rceil] \text{ bits/vecteur} \tag{2.15}$$

La partie codage sans perte du schéma a été étudiée dans le cadre de la thèse de Teddy Voinson [111], qui a notamment adapté aux énergies un codeur par plage dérivé du codage run-length (codeur STACK RUN [109]). Ce type de codage permet de tirer parti des redondances résiduelles issues de la quantification. Il permet d'atteindre des débits se situant en deçà de la borne inférieure de l'entropie du premier ordre de la formule (2.13). Néanmoins, dans ce mémoire nous nous bornerons à baser nos résultats sur la formule du débit (2.15).

Les performances d'un algorithme de compression basé sur la QVA sont donc directement liées à la distribution de l'énergie des vecteurs d'ondelettes. En effet, sa statistique est exploitée d'une part par un algorithme de compression sans perte, et d'autre part, la population des couches du dictionnaire dépend du rayon. Cette distribution est liée à la forme des vecteurs suivant la manière de prendre en compte les dépendances spatiales des coefficients d'ondelettes de la TO3D.

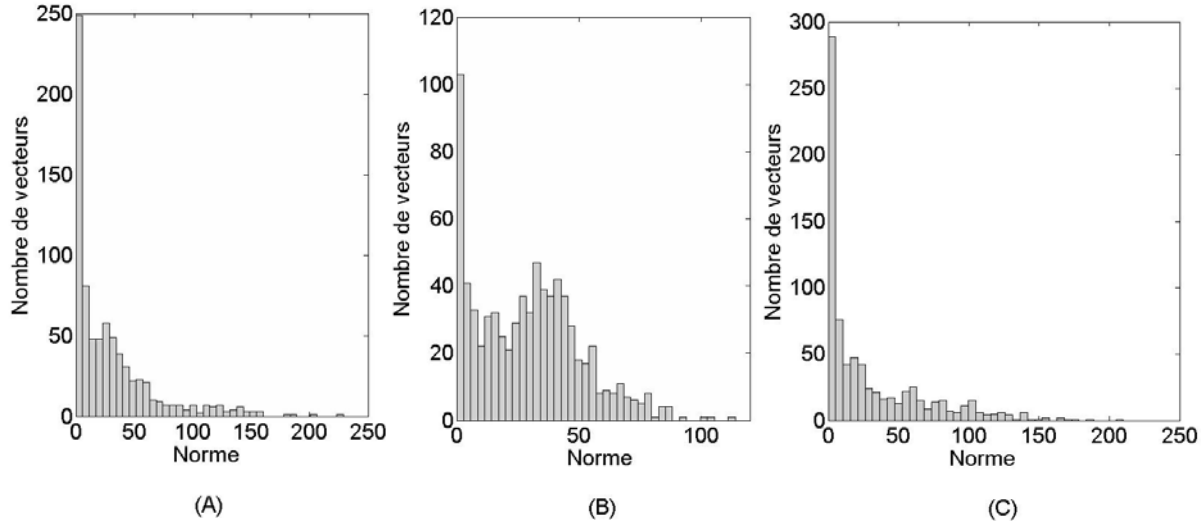

Fig. 2.8: Comparaison des histogrammes des normes L^1 avec des vecteurs orientés verticalement (A), horizontalement (B) et cubiquement (C) - sous-bande LHL3 de MR2.

2.3 Orientation des vecteurs

Le schéma de QVA employé utilise un dictionnaire pyramidal, entrainant un indexage basé sur la norme L^1. Cette norme[3] peut être vue comme la magnitude moyenne des composantes d'un vecteur. La conséquence d'une telle remarque est la suivante : la norme L^1 constitue une bonne mesure de l'activité d'un voisinage de l'image à une certaine échelle. Nous allons à présent étudier de quelle manière les dépendances spatiales influent sur l'énergie (au sens de la norme L^1) des vecteurs des sous-bandes de la TO3D. Pour discuter de ce problème, la figure 2.8 montre une comparaison entre les histogrammes de la norme de trois formes de vecteurs sur la sous-bande LHL3 (détails verticaux + filtrage passe-bas le long de l'axe z) de MR2 de l'étude E1 :

A Vecteurs verticaux $(8 \times 1 \times 1)$

B Vecteurs horizontaux $(1 \times 8 \times 1)$

C Vecteurs cubiques $(2 \times 2 \times 2)$

La distribution A est bien plus piquée que B : des vecteurs avec la même orientation que les détails d'ondelettes (orthogonale à la direction du filtre) capturent bien mieux le phénomène d'agglutinement[4] des coefficients significatifs et, inversement augmentent la concentration autour de zéro. Ce point montre que dans un souci d'efficacité de codage, il ne faut surtout pas privilégier les vecteurs orthogonaux à la direction des sous-bandes 2D.

De plus, ayant effectué une TO3D, il existe une corrélation supplémentaire le long de la direction temporelle (z) comme nous pouvons le constater sur la distribution C, avec des vecteurs possédant une orientation cubique où il y a encore plus de concentration de l'énergie en zéro.

Les courbes débit-distorsion des trois orientations confirment également que des vecteurs offrant une statistique plus piquée minimisent la distorsion. Les différences entre les histogrammes se répercutent bien sur les courbes R-D (figure 2.9). En effet, plus la distribution de la norme est piquée, meilleur est le codage entropique de celle-ci (figure 2.7). Ainsi une orientation cubique est un bon compromis pour capturer l'agglutinement dans les trois directions de toutes les sous-bandes

[3] Rappel : Pour un vecteur X de taille n, on a $\|X\|_1 = \sum_{i=1}^{n} |x_i|$

[4] Clustering en anglais

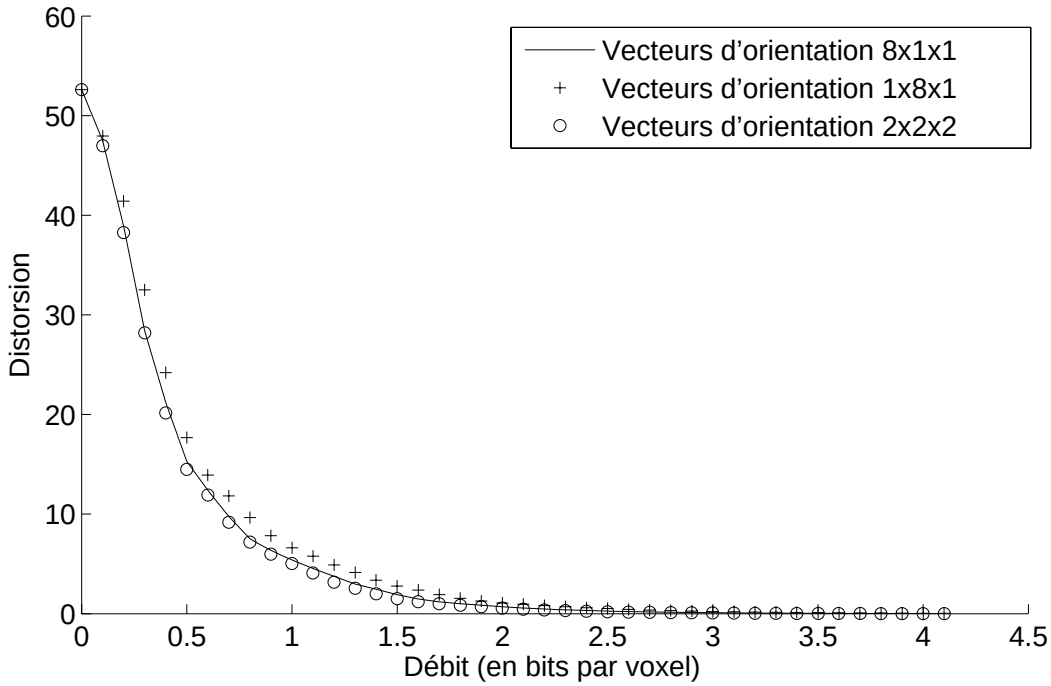

Fig. 2.9: Comparaison des courbes débit-distorsion obtenues par QVA (Norme L^1) avec des vecteurs orientés verticalement, horizontalement et cubiquement (C) - sous-bande LHL3 de MR2.

conduisant à une statistique avantageuse. Cette forme de vecteur a donc été choisie dans l'ensemble de nos travaux. L'incidence de l'orientation des vecteurs sur la statistique de l'énergie montre l'importance des dépendances spatiales dans le domaine multirésolution. Cela va donc à l'encontre de l'hypothèse classique i.i.d pour modéliser les coefficients d'ondelettes. En fait, l'hypothèse i.i.d est valable sur les petites sous-bandes ou si l'on prend des vecteurs dont les composantes ne sont pas voisines. Dans le chapitre 4, nous proposerons une modélisation qui permet de s'affranchir de cette hypothèse.

L'étude des caractéristiques des blocs d'ondelettes aboutit à la notion de vecteurs significatifs relatifs à l'état de faible énergie. Nous avons en outre relevé l'importance de l'orientation des vecteurs : la forme des vecteurs permet donc d'adapter le quantificateur aux structures présentes dans la sous-image d'ondelettes. Nous introduisons ici un nouveau quantificateur pour exploiter cette très forte concentration de vecteurs de faible énergie des sous-bandes de la TO3D pour une image médicale. Il s'appuie sur une technique de seuillage vectoriel que nous avons définie par le concept de Zone Morte vectorielle (ZM).

2.4 Zone Morte Vectorielle en imagerie médicale

Dans le cas scalaire, la notion de coefficient d'ondelette significatif a abouti à l'utilisation d'une zone morte dans les schémas de compression les plus récents (comme par exemple JPEG2000 [107]), permettant d'améliorer le compromis débit-distorsion. Le principe consiste à augmenter la redondance des coefficients nuls en élargissant la première cellule de quantification, les coefficients significatifs sont ainsi quantifiés plus finement et le compromis débit-distorsion est amélioré.

Cette propriété a été étendue au cas vectoriel [110],[111], en généralisant à la quantification

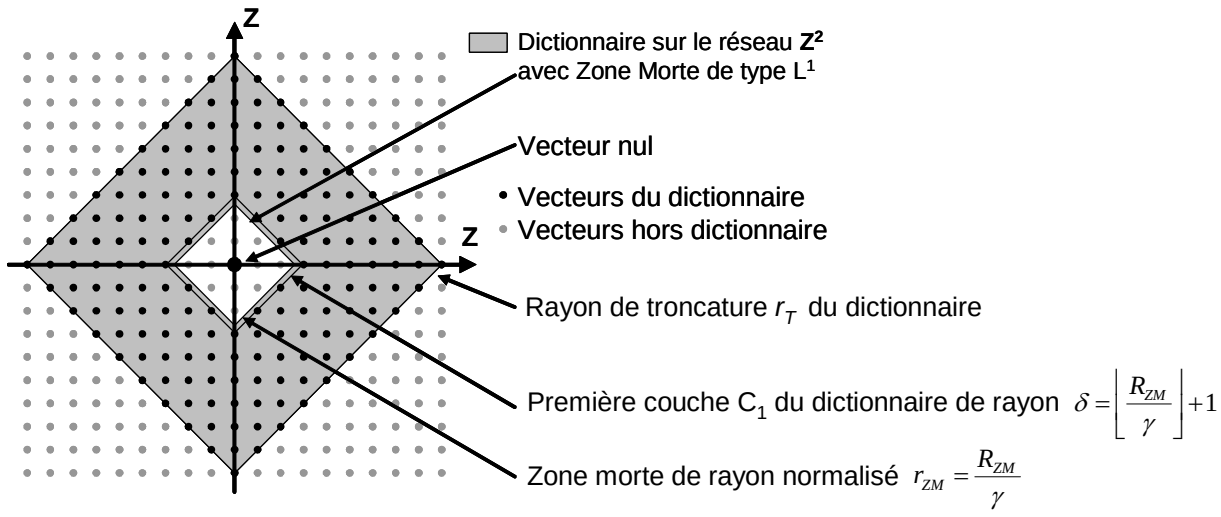

FIG. 2.10: Dictionnaire avec zone morte pyramidale sur le réseau $\mathbb{Z}^2$.

vectorielle algébrique le principe de la zone morte, à savoir, en élargissant la cellule de Voronoï du vecteur nul mais également en modifiant sa forme. La zone morte vectorielle permet de diminuer le débit entropique de l'énergie en remplaçant par zéro les vecteurs dont l'énergie est en deçà d'un seuil correspondant au rayon de la zone morte. Nous montrerons qu'elle peut être efficacement appliquée aux images médicales dans le domaine des ondelettes 3D qui contiennent de grandes zones non significatives.

Ce paragraphe comporte deux parties, la première présente la zone morte vectorielle pour un dictionnaire de forme quelconque. La seconde donne les détails de la QVAZM 3D dans le cas d'un dictionnaire pyramidal.

2.4.1 Définition de la zone morte vectorielle et du dictionnaire associé

Soit un réseau Λ de dimension n. On désigne par zone morte vectorielle ZM la cellule de Voronoï déformée contenant le vecteur nul, de rayon $R_{ZM} \in \mathbb{R}^+$ définie par [110] :

$$ZM = S_d^n(R_{ZM}) = \{X \in \mathbb{R}^n / d(0, X) \leq R_{ZM}\} \tag{2.16}$$

où d est une distance déterminant la forme de la zone morte vectorielle.

La zone morte vectorielle a une forme différente des autres cellules de Voronoï : tandis que les autres sont définies selon le critère (2.2) par la distance euclidienne (norme L^2), la zone morte se définit par rapport à la distance d. La figure 2.10 représente la ZM dans le cas où la distance d est la norme L^1.

Soit γ le facteur de projection (ou de mise à l'échelle) appliqué à la source. On note δ le rayon de la première couche C_1 hors zone morte :

$$\delta = \left\lfloor \frac{R_{ZM}}{\gamma} \right\rfloor + 1 \tag{2.17}$$

avec $\lfloor . \rfloor$ la partie entière.

Le dictionnaire de rayon de troncature r_T obtenu à partir d'un réseau Λ avec zone morte vectorielle est défini par :

$$C_{ZM} = \{Y \in \Lambda / \delta \leq d(0, Y) \leq r_T\} \cup \{\mathbf{0}\} \tag{2.18}$$

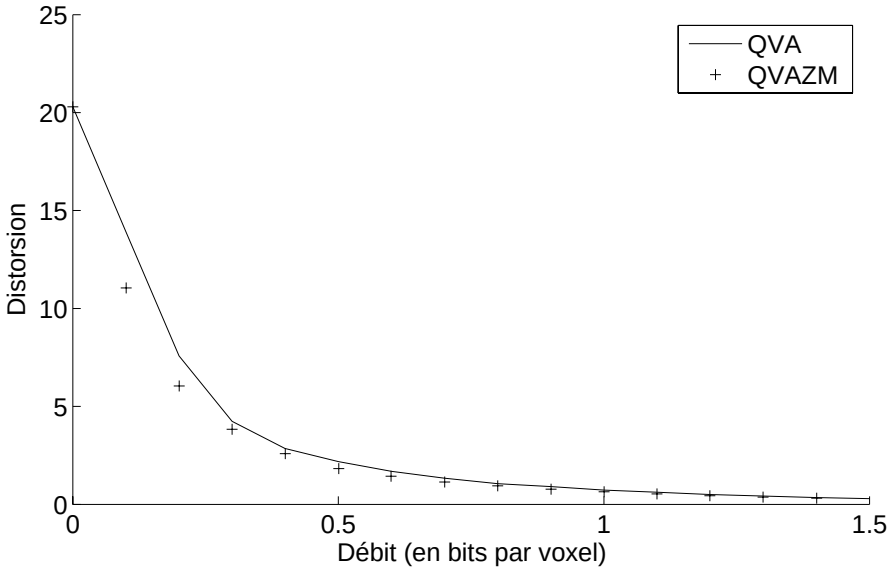

Fig. 2.11: Comparaison des courbes débit-distorsion obtenues par QVA, par QVAZM - sous-bande LLH3 de MR4.

avec $\mathbf{0}$ est le vecteur nul.

La figure 2.10 représente (en gris) un tel dictionnaire dans le réseau $\mathbb{Z}^2$ dans le cas de la norme L^1. Le réseau construit à partir de la formule (2.18) admet une différence majeure par rapport à ceux décrits dans l'état de l'art : la plupart des travaux ont cherché à déterminer, pour différentes dimensions, les réseaux optimaux au sens de la théorie haute résolution. Jeong et *al.* [53] ont certes suggéré l'utilisation de réseaux déformés, mais à l'image de la quantification optimale de Max-Loyd, cette méthode privilégie les vecteurs avec une très grande probabilité d'apparition au détriment des autres.

Dans notre cas, il s'agit d'exploiter les propriétés des blocs d'ondelettes mis en évidence dans le chapitre précédent : modifier la forme et la taille de la cellule de Voronoï d'origine a pour but d'améliorer les performances des algorithmes de compression sans perte, tels que les codeurs entropiques ou stack run, employés pour coder les énergies dans le code produit CP défini dans (2.12). Nous avons montré que la distribution de l'énergie des vecteurs admettait un pic proche de zéro. Ce pic est une conséquence de la mémoire du signal : les coefficients de forte énergie ont tendance à s'agglutiner, et, de la même manière, un coefficient de faible valeur a une forte probabilité d'appartenir à un bloc de faible énergie. Par conséquent l'utilisation d'un seuillage sur la norme permet, non seulement d'améliorer le compromis débit distorsion, mais également de mieux coder les vecteurs significatifs au détriment des vecteurs non significatifs. La figure 2.11 compare la QVA 3D et la QVAZM 3D en terme de performance débit/distorsion sur la sous-bande LLH3 de MR4. Pour un débit inférieur à 0,5 bits/voxel, elle illustre le gain important en compromis débit/distorsion obtenu en utilisant une zone morte vectorielle.

Le débit DB_{ZM} de la source quantifiée se déduit de (2.15) :

$$DB_{ZM} = -P(0)\log_2(P(0)) - \sum_{r=\delta}^{r_T} P(r)\left[\log_2(P(r)) - \lceil \log_2(N(r)) \rceil\right] \text{ bits/vecteur} \qquad (2.19)$$

avec $P(0) = P(X \in ZM)$ la probabilité d'appartenance à la zone morte.

Le débit donné par (2.19) varie en fonction du rayon R_{ZM} de la zone morte vectorielle et du facteur de projection γ : il y a donc deux paramètres à déterminer pour régler le débit de la source quantifié. Ce point sera explicité dans le paragraphe 2.5.

Nous allons à présent nous attacher à détailler la procédure de quantification sur $\mathbb{Z}^n$ dans le cas d'une zone morte pyramidale.

2.4.2 Zone morte pyramidale sur le réseau $\mathbb{Z}^n$

D'après les formules (2.16) et (2.18), on définit la zone morte pyramidale et le dictionnaire associé de la manière suivante :

Zone morte vectorielle pyramidale

$$ZM = \{X \in \mathbb{R}^n / \|X\|_1 \leq R_{ZM}\} \tag{2.20}$$

Dictionnaire avec zone morte pyramidale

$$C_{ZM} = \{Y \in \mathbb{Z}^n / \delta \leq \|Y\|_1 \leq r_T\} \cup \{\mathbf{0}\} \tag{2.21}$$

Dans la suite, on notera r_{ZM} le rayon de la zone morte après projection de la source et défini par :

$$r_{ZM} = \frac{R_{ZM}}{\gamma} \tag{2.22}$$

Le dictionnaire défini par (2.21) est composé de trois zones correspondant à trois types de quantification : la zone morte ZM, la zone de surcharge ZS et la zone de quantification uniforme ZU. Ces trois zones sont représentées sur la figure 2.12.

Les vecteurs appartenant à la zone morte sont quantifiés par le vecteur nul. Les vecteurs de ZU sont quantifiés selon la règle de quantification du réseau employé (dans le cas du réseau $\mathbb{Z}^n$ les composantes des vecteurs sont quantifiées par leur entier le plus proche). Cette zone est censée contenir les vecteurs significatifs de la source. L'hypothèse haute résolution est valide dans cette région. La distorsion de surcharge engendrée par la quantification des vecteurs hors du dictionnaire n'est pas prise en compte : nous supposons que le rayon de troncature est assez grand pour englober toute la source.

La zone de surcharge qui constitue l'interface entre ZM et ZU est définie par :

$$ZS = \left\{\bar{X} \in \mathbb{R}^n / \left\|Q\left(\bar{X}\right)\right\|_1 < \delta \text{ et } \left\|\bar{X}\right\|_1 > r_{ZM}\right\} \tag{2.23}$$

ZS correspond à l'ensemble des vecteurs, hors de la zone morte, qui ne sont pas naturellement quantifiés en dehors de la zone morte (voir figure 2.12) : ces vecteurs doivent donc faire l'objet d'une procédure particulière pour être quantifiés par des vecteurs du dictionnaire de la zone uniforme. Plusieurs méthodes ont été proposées pour limiter la distorsion de surcharge induite par cette opération. La première consiste en une reprojection itérative de ces vecteurs par un facteur de type $1 + \varepsilon$ avec ($\varepsilon << 1$) jusqu'à ce qu'ils sortent de la zone de surcharge [111]. Une autre méthode de type projection orthogonale est proposée dans [46]. Nous ne détaillerons pas plus cette zone de surcharge. En effet, expérimentalement elle contribue très peu au débit et à la distorsion.

Pour résumer, après l'étape de TO3D, notre schéma de compression découpe les vecteurs suivant une orientation cubique ($2 \times 2 \times 2$) pour toutes les sous-bandes. Les vecteurs découpés sont ensuite quantifiés en utilisant une QVA sur le réseau $\mathbb{Z}^n$ auquel on a choisit d'associer la norme L^1. De plus, nous avons introduit dans le processus de quantification une zone morte vectorielle pour minimiser le compromis débit-distorsion. Une fois les vecteurs quantifiées, le codage est réalisé par le biais

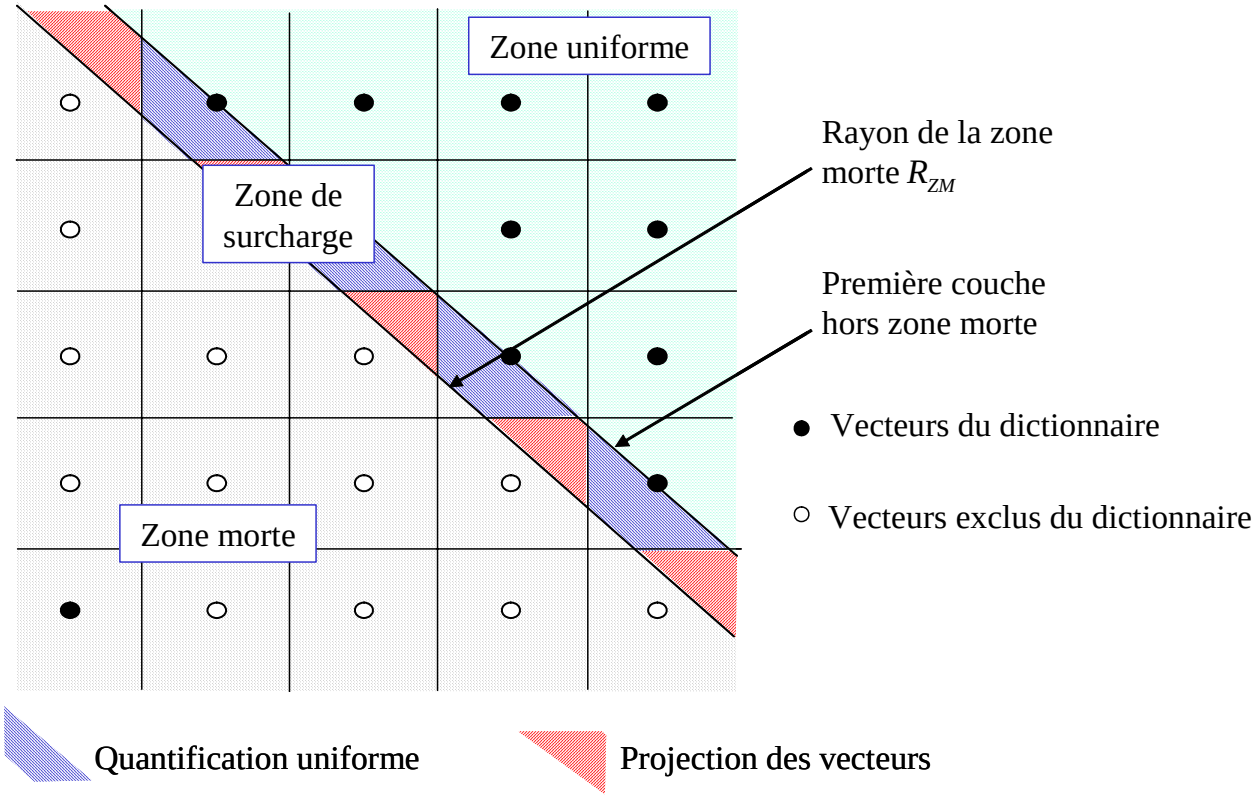

Fig. 2.12: Les trois zones de quantification pour le réseau $\mathbb{Z}^2$ doté d'une zone morte de pyramidale.

d'un code produit séparant le codage de la norme et de la position (figure 2.7). Cependant, pour que le schéma s'inscrive dans une chaîne complète de compression, nous devons régler les débits de chaque sous-bande dans une procédure d'allocation de débits correspondant à un problème classique de minimisation sous contrainte d'égalité. Dans le cas de la QVAZM 3D, cette minimisation nécessite une étape d'optimisation supplémentaire par rapport à un schéma classique : déterminer le paramètre de seuillage optimal (le rayon de la zone morte). Le paragraphe suivant expliquera comment régler ce paramètre par le biais d'un algorithme de recherche rapide et s'attachera ensuite à présenter l'allocation de débits. Cette étape qui est un point crucial de la faisabilité de notre chaîne de compression peut être une étape extrêmement coûteuse pour les gros volumes de données médicales. Cependant, dans ce chapitre, nous résoudrons cette optimisation sous contraintes par des méthodes numériques classiques assurant une bonne précision au détriment de la complexité. Des méthodes basées sur l'approximation des courbes débit-distorsion et sur des modèles statistiques seront proposées dans le dernier chapitre. Certaines pourront évidemment être appliquées aux images médicales volumiques.

2.5 Allocation de débits

Ce paragraphe est consacré à la présentation d'une méthode d'allocation des débits associée à la QVAZM 3D. Cet algorithme détermine les débits optimaux de chaque sous-bande produite par la TO3D qui minimisent la distorsion totale avec pour contrainte d'atteindre le débit cible, c'est à dire le taux de compression souhaité.

De manière très synthétique, le principe général de l'optimisation numérique peut se décomposer de la manière suivante :

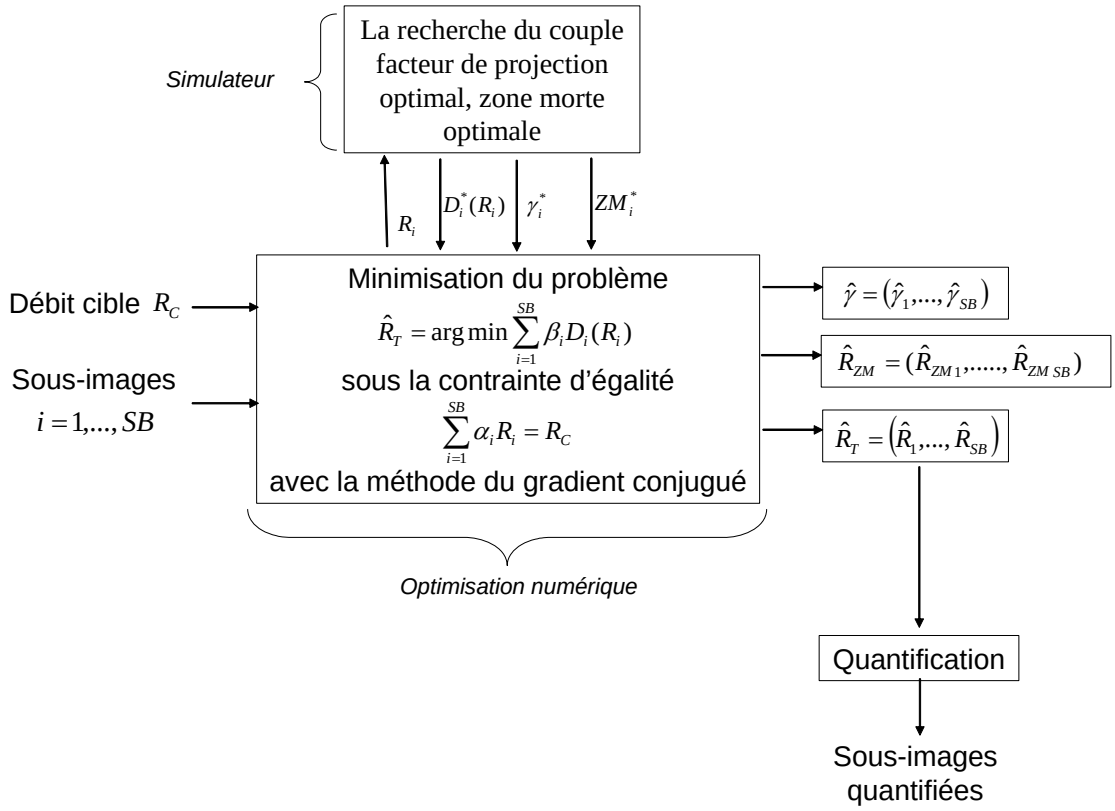

FIG. 2.13: Allocation des débits de la QVAZM 3D par une méthode lagrangienne utilisant comme simulateur l'algorithme de recherche du couple facteur de projection optimal/zone morte optimale.

Simulateur Il est chargé de collecter les données propres au problème de minimisation. Dans notre cas ce simulateur se compose des SB fonctions distorsion-débit $R_i \mapsto D_i(R_i)$, i étant l'indice de la sous-bande. Ces fonctions sont obtenues par l'algorithme rapide de recherche de la zone morte optimale décrit par la suite.

Optimisation Un algorithme d'optimisation faisant appel au simulateur se charge de converger vers la solution. C'est l'algorithme de type lagrangien qui est résumé dans le sous-paragraphe 2.5.2.

Il est crucial de différencier la mise au point du simulateur de celle de l'algorithme d'optimisation : le premier est spécifique au schéma de compression tandis que le deuxième est une méthode mathématique adaptée. L'algorithme d'optimisation peut donc être utilisé pour différents simulateurs : l'algorithme proposé dans [86] a par exemple été employé avec un simulateur obtenu par quantification scalaire [87] et par QVA [89] .

Au final, en couplant le simulateur et l'optimisation, l'objectif de l'allocation optimale des débits avec la QVAZM 3D consiste à trouver les facteurs de projection optimaux $\widehat{\gamma} = (\widehat{\gamma_1}, ..., \widehat{\gamma_{SB}})$ et les rayons de zones mortes optimaux $\widehat{R_{ZM}} = \left(\widehat{R_{ZM_1}}, ..., \widehat{R_{ZM_{SB}}}\right)$ pour les SB sous-bande de la TO3D. La figure 2.13 résume l'allocation de débit utilisée pour la QVAZM 3D.

2.5.1 La recherche du couple facteur de projection optimal, zone morte optimale

Nous allons nous attacher à résoudre le problème de recherche de la zone morte optimale, c'est-à-dire à trouver le rayon de la zone morte qui, pour un débit cible donné, minimise la distorsion.

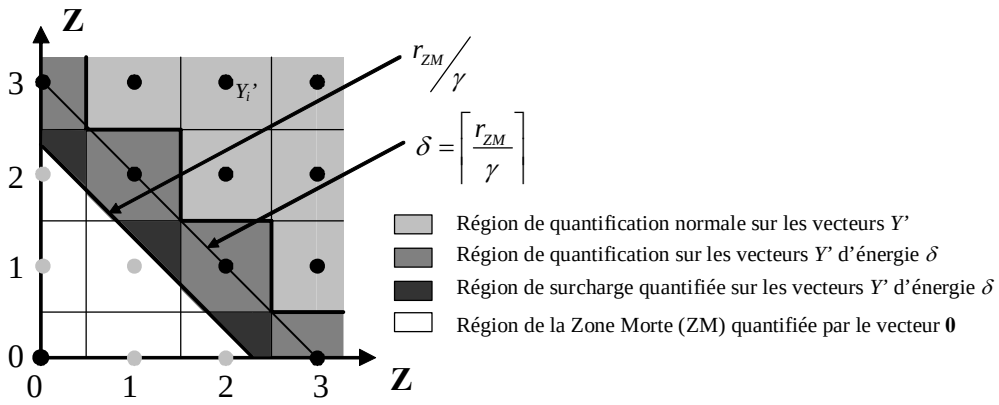

Fig. 2.14: Représentation des trois zones de quantification après projection de la source.

C'est un point crucial de la faisabilité de la chaîne globale d'allocation des débits. Nous allons montrer que ce problème peut être reformulé par un problème de minimisation d'une fonction à une seule variable : la distorsion en fonction du rayon de la zone morte. Nous réaliserons une étude des caractéristiques de cette fonction pour une sous-bande produite par une TO3D. Cela conduira à l'utilisation d'un algorithme quasi-optimal dont le coût de calcul est très inférieur à celui de l'algorithme optimal.

2.5.1.1 Reformulation du problème

On se place ici dans une sous-bande donnée, pour laquelle le débit à atteindre est R_C. La recherche de la zone morte optimale est un problème de minimisation sous contrainte d'égalité. En effet, pour un débit R_C à atteindre, le couple optimal rayon/facteur de projection, noté $\left(R_{opt}, \gamma_{opt}\right)$, vérifie la relation suivante :

$$\begin{aligned} \left(R_{opt}, \gamma_{opt}\right) &= \arg \min_{(R_{ZM},\gamma)} \left\{D\left(R_{ZM}, \gamma\right)\right\}, \\ \text{sous la contrainte } R\left(R_{ZM}, \gamma\right) &= R_C \end{aligned} \tag{2.24}$$

avec $D\left(.,.\right)$ la distorsion de la source quantifiée et $R\left(.,.\right)$ le débit.

Les définitions ci-dessous ont pour objectif de simplifier le problème (2.24).

Dans la suite, la zone morte sera exprimée en terme de rayon normalisé r_{ZM} (voir figure 2.14) dont nous rappelons la définition :

$$r_{ZM} = \frac{R_{ZM}}{\gamma} \tag{2.25}$$

avec R_{ZM} le rayon de la zone morte et γ le facteur de projection employé.

Redéfinissons tout d'abord les deux fonctions débit et distorsion du problème (2.24) :

Définition 1 : Débit

$$\begin{aligned} R\colon \mathbb{R}^+ \times \mathbb{R}^+ &\rightarrow \mathbb{R}^+ \\ \left(r_{ZM}, \gamma\right) &\mapsto R\left(r_{ZM}, \gamma\right) \triangleq R\left(\gamma r_{ZM}, \gamma\right) \end{aligned}$$

Définition 2 : Distorsion

$$\begin{aligned} D\colon \mathbb{R}^+ \times \mathbb{R}^+ &\to \mathbb{R}^+ \\ (r_{ZM}, \gamma) &\mapsto D(r_{ZM}, \gamma) \triangleq D(\gamma r_{ZM}, \gamma) \end{aligned}$$

Ces deux grandeurs varient en fonction du rayon de la zone morte et du facteur de projection. Le problème (2.24) se réécrit alors :

$$\begin{aligned} (r_{opt}, \gamma_{opt}) &= \arg \min_{(r_{ZM},\gamma)} \{D(r_{ZM}, \gamma)\}, \qquad (2.26) \\ \text{sous la contrainte } R(r_{ZM}, \gamma) &= R_C \end{aligned}$$

avec R_C le débit cible.

La deuxième partie de la reformulation consiste à transformer le problème de minimisation à deux variables (2.26) en un problème de minimisation d'une fonction unidimensionelle en considérant les deux fonctions suivantes :

Définition 3 On note γ^* la fonction qui, pour un rayon donné, associe le facteur de projection permettant d'atteindre le débit cible :

$$\gamma^*\colon \mathbb{R}^+ \to \mathbb{R}^+ \qquad (2.27)$$

$$r_{ZM} \mapsto \gamma^*(r_{ZM}), \qquad (2.28)$$

$$\text{tel que } R(r_{ZM}, \gamma^*(r_{ZM})) = R_C$$

La fonction $R : \gamma \to R(r_{ZM}, \gamma)$ est monotone décroissante : à mesure que la source se concentre sur un nombre de couches du dictionnaire de plus en plus petit, l'entropie de la source quantifiée diminue. L'existence et l'unicité de $\gamma^*(r_{ZM})$ est donc toujours vérifiée. $\gamma^*(r_{ZM})$ s'obtient en résolvant l'équation $R(r_{ZM}, \gamma^*(r_{ZM})) - R_C = 0$ par une méthode classique de recherche du zéro d'une fonction [84].

Définition 4 On note D^* la fonction définie par :

$$\begin{aligned} D^*\colon \mathbb{R}^+ &\to \mathbb{R}^+ \qquad (2.29) \\ r_{ZM} &\mapsto D^*(r_{ZM}) = D(r_{ZM}, \gamma^*(r_{ZM})) \end{aligned}$$

Le problème de recherche de la zone morte optimale se formalise alors de la manière suivante :

$$\begin{aligned} r_{opt} &= \arg \min_{r_{ZM}} \{D^*(r_{ZM})\} \qquad (2.30) \\ \gamma_{opt} &= \gamma^*(r_{opt}) \end{aligned}$$

Nous venons de passer d'un problème de minimisation sous contrainte d'égalité à une simple minimisation d'une fonction à une variable D^*.

2.5.1.2 Courbe débit-distorsion en fonction de la zone morte, et zone morte optimale

La figure 2.15 représente un exemple de la fonction D^*. La minimisation du problème (2.30), repose donc sur les propriétés de cette fonction. Deux propriétés sont discutées ici, la première va permettre de restreindre l'intervalle de recherche en considérant le volume de la zone morte vectorielle. La seconde traite des discontinuités de la fonction D^* qu'il est impératif de localiser dans l'optique de l'utilisation de techniques classiques de minimisation de fonctions continues.

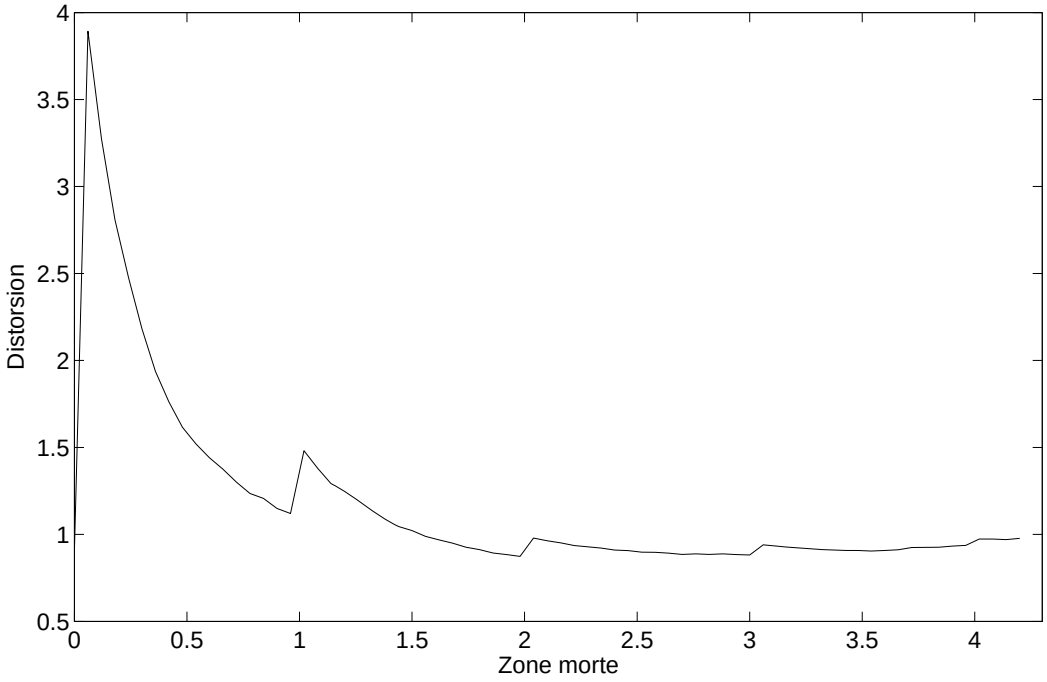

Fig. 2.15: Courbe distorsion en fonction du rayon de la zone morte à un débit cible de 1 bit/voxel - sous-bande LHL3 de CT1.

Intervalle de recherche : Comme on peut le constater sur la figure 2.15, la fonction D^* admet un maximum proche de zéro qui s'explique du point de vue de la quantification : pour un rayon de zone morte inférieur à 1, la zone morte n'englobe aucune couche du dictionnaire. Le gain en terme d'entropie est alors limité. Ainsi pour un même débit, l'erreur sera plus grande augmentant la fonction D^*. Par ailleurs, lorsque le rayon tend vers zéro, la zone morte s'inscrit dans la cellule de Voronoï cubique du schéma classique de QVA : on obtient un effet inverse à celui escompté, les vecteurs de faible énergie sont projetés hors de la zone morte. La distorsion supplémentaire engendrée par les nombreux vecteurs ressortis artificiellement du Voronoï d'origine provoque le pic en 0^+ qui apparaît sur la figure 2.15.

En conséquence, dans la majeure partie des cas, il est peu probable d'obtenir une zone morte optimale de rayon inférieure à 1. La recherche s'effectuera donc sur l'intervalle $]1, +\infty[$.

Discontinuités de D^* : La fonction D^* admet des discontinuités en l^-, avec $l \in \mathbb{N}$. Ces discontinuités proviennent du phénomène de saut de couche : le volume de la zone de surcharge tend à devenir minimal lorsque le rayon de la zone morte tend vers l^-. Au contraire, lorsque le rayon tend vers l^+, le volume de la zone de surcharge devient maximal. Or, la projection des vecteurs appartenant à ZS induit une distorsion supplémentaire. Cet accroissement s'effectue sans gain au niveau du débit puisque ces vecteurs se retrouvent sur une surface plus peuplée.

En prenant en compte les deux points précédents, la recherche du minimum de la fonction D^* va consister à effectuer une recherche dichotomique sur les intervalle de type $I_l = [l, l+1[$. Par conséquent, l'algorithme optimal consiste à appliquer une méthode de recherche du minimum d'une fonction à une variable [84] sur les intervalles I_l, avec $l = 1, ..., L$.

Nous venons de transformer la recherche de la zone morte optimale en une application d'une technique de minimisation d'une fonction à une variable par intervalle. Cependant, cette approche est susceptible de ne pas respecter la contrainte de complexité. En effet, le calcul de D^* nécessite de

nombreux appels aux données. Ce point peut s'avérer dramatique lorsque la taille des sous-bandes est importante, ce qui est le cas des images médicales 3D. Il apparaît donc crucial de proposer une méthode de réglage de la zone morte vectorielle limitant le nombre de calculs effectués à partir des données.

2.5.1.3 Algorithme quasi-optimal

Nous allons maintenant présenter un algorithme rapide basé sur différentes propriétés de la courbe D^* [46]. La recherche de minimum par morceau peut être remplacée par une simple procédure itérative sur un ensemble discret de zone morte décrit par la suite (2.32). De plus, les minimas locaux de chaque intervalle I_l se trouvent sur une enveloppe convexe fournissant ainsi un critère d'arrêt. Enfin, un encadrement simple (2.36) permet de réduire à chaque itération l'intervalle de recherche du facteur de projection correspondant à la zone morte testée.

Décroissance par Morceaux : La quantification dans la zone de surcharge engendre une distorsion supérieure à celle de la quantification uniforme. Ce point a déjà été évoqué dans le paragraphe précédent : les vecteurs de la zone de surcharge doivent être projetés sur le réseau de manière spécifique. De plus, cette quantification augmente le débit binaire du code suffixe puisque les vecteurs sont projetés sur une couche de plus grand rayon. Ce surcoût a une influence sur la valeur de D^* lorsque le nombre de vecteurs concernés devient important. Il y a donc décroissance de la fonction D^* sur les intervalles I_l : la distorsion diminue grâce à l'effet croisé de l'augmentation de la zone morte et de la diminution de la zone de surcharge.

La recherche du minimum de l'intervalle I_l devient dès lors :

$$\begin{aligned} r_{opt} &= \underset{r_{ZM}\in\{l-\epsilon, l\in\mathbb{N}\}}{\arg\min} \{D^*(r_{ZM})\} \\ \gamma_{opt} &= \gamma^*(r_{opt}) \end{aligned} \tag{2.31}$$

avec $\epsilon << 1$ la constante fixant *a priori* la position du minimum local.
La recherche de la zone morte devient une séquence de test de valeurs de type :

$$r_l = l - \varepsilon \tag{2.32}$$

C'est le point central de la méthode rapide que nous proposons ici, car le domaine de recherche se restreint à un ensemble dénombrable ce qui conduit à une diminution notable de la complexité : l'étape de recherche du minimum de la fonction D^* sur l'intervalle I_l est remplacée par le calcul de $\gamma^*(.)$ pour quelques valeurs test.

Convexité : Les minima locaux de chaque intervalle I_l se trouvent sur une enveloppe convexe. Cette propriété est toujours vérifiée lorsque le gain apporté par la zone morte est significatif (voir figure 2.15). Par conséquent, $D^*(r_{l+1}) > D^*(r_l)$ implique que r_l est le rayon recherché.

La recherche de la zone morte peut encore être accélérée en étudiant l'intervalle de recherche de $\gamma^*(.)$.

Encadrement du facteur de projection :

Le calcul de la fonction $\gamma^*(.)$, c'est-à-dire l'obtention du facteur de projection permettant d'atteindre le débit cible pour une zone morte donnée, nécessite l'emploi de techniques de recherche de zéro d'une fonction, également coûteuses du fait de l'appel aux données. Nous proposons de diminuer le nombre de calculs en encadrant le domaine de recherche du zéro.

Soient R^1_{ZM}, $R^2_{ZM} \in \mathbb{R}^+$ deux rayons de zone morte tels que $R^1_{ZM} < R^2_{ZM}$, on note γ^*_1 et γ^*_2 les facteurs de projection correspondants, permettant d'atteindre le débit cible R_C. On a

$$\gamma^*_1 > \gamma^*_2 \tag{2.33}$$

Il s'agit là du principe même de la zone morte : pour un débit cible donné, plus le rayon de la zone morte est grand, plus les vecteurs hors zone morte sont quantifiés finement et donc plus le facteur de projection est petit.

Il en est de même si l'on exprime la zone morte en nombre de couches correspondant à la zone morte projetée :

Soient r_1, $r_2 \in \mathbb{R}^+$ tels que $r_1 < r_2$ alors,

$$\gamma^*(r_1) > \gamma^*(r_2) \tag{2.34}$$

Posons maintenant $R^1_{ZM} = r_1\gamma^*_1$ et $R^2_{ZM} = r_2\gamma^*_2$, avec $\gamma^*_1 = \gamma^*\left(R^1_{ZM}\right)$ et $\gamma^*_2 = \gamma^*\left(R^2_{ZM}\right)$.
En supposant que $R^1_{ZM} < R^2_{ZM}$, on a : $r_1\gamma^*_1 < r_2\gamma^*_2 \Leftrightarrow \frac{r_1}{r_2}\gamma^*_1 < \gamma^*_2$
D'où l'encadrement suivant :

$$\frac{r_1}{r_2}\gamma^*_1 < \gamma^*_2 < \gamma^*_1 \tag{2.35}$$

Cette relation va permettre de réduire de manière significative le domaine de recherche de γ^*. En effet, si l'on considère la suite $(r_l)_{l\in N}$ définie plus haut, on obtient l'encadrement suivant :

$$\frac{r_l}{r_{l+1}}\gamma^*(r_l) < \gamma^*(r_{l+1}) < \gamma^*(r_l) \tag{2.36}$$

avec $r_l = l - \varepsilon \ \forall l \in \mathbb{N}$.

Il est également possible d'ajouter d'autres encadrements qui seront détaillés dans le chapitre 4. Ils sont basés sur une logique d'importance des sous-bandes dans le schéma complet. Ainsi, pratiquement, quatre ou cinq itérations suffisent à trouver γ^*.

Nous allons maintenant présenter l'algorithme rapide basé sur les 3 propriétés décrites ci-dessus. L'algorithme proposé est sous-optimal : il n'assure pas de trouver le minimum global en raison du choix arbitraire de l'ensemble des valeurs test r_k. Toutefois, les tests ont montré qu'un choix du type $\epsilon = 0,1$ permet d'obtenir des résultats presque identiques à ceux correspondant à la zone morte optimale. La figure 2.16 confirme ce résultat et montre que l'algorithme rapide permet d'améliorer la distorsion par rapport au schéma QVA.

Algorithme rapide (quasi optimal) de recherche de la zone morte

Algorithme 1 *Début*
Initialisation de ε
Initialisation du débit cible R_C
Initialisation de la zone morte : $r_0 = 0$
Calcul de $D^*_0 = D^*(r_0)$ *et de* γ^*_0
$arrêt = 0$
$l = 0$
Tant que $(arrêt = 0)$
$l = l + 1$
$r_l = l - \varepsilon$
Recherche de $\gamma^*_l(r_l)$ *sur* $\left[\frac{r_{l-1}}{r_l}\gamma^*_{l-1}\ ,\ \gamma^*_{l-1}\right]$
Calcul de $D^*_l = D^*(r_l)$
Si $\left(D^*_l > D^*_{l-1}\right)$
$r_{opt} = r_{l-1}$
$\gamma_{opt} = \gamma^*_l$
$arrêt = 1$
Fin si
Fin tant que
Fin

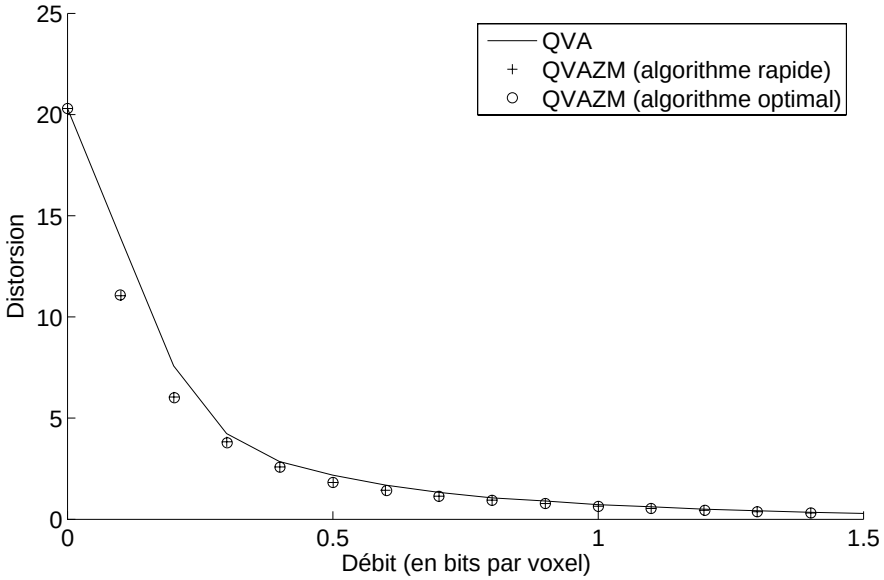

FIG. 2.16: Comparaison des courbes débit-distorsion obtenues par QVA, par QVAZM avec recherche optimale des paramètres, et par QVAZM avec l'algorithme rapide de recherche de la zone morte - sous-bande LLH3 de MR4.

Nous venons de présenter un algorithme permettant de contrôler l'accroissement de complexité du schéma de QVAZM par rapport au schéma de QVA classique. Cet algorithme chargé de collecter les données est utilisé au niveau supérieur du processus de compression, à savoir l'allocation des débits présentée dans le paragraphe suivant.

2.5.2 Optimisation numérique classique associée à la QVAZM 3D

Le schéma de compression par QVAZM 3D, proposé dans cette thèse appartient à l'approche intra-bande. Ainsi chaque sous-bande est codée indépendamment. Pour atteindre un débit cible donné, une approche intra-bande doit procéder à une allocation qui consiste à déterminer les débits à attribuer à chaque sous-bande.

De manière générale, l'allocation optimale de débit consiste à répartir le débit cible R_C sur chaque sous-bande i issue de la TO3D afin de minimiser la distorsion moyenne totale pondérée D_T. Le problème d'optimisation peut s'écrire comme un problème classique de minimisation multidimensionnel avec contrainte d'égalité :

$$\begin{cases} \min\limits_{\{R_i \in \mathbb{R} | i=1,...,SB\}} D_T\left(R_{LLL_D}, R_1, ..., R_i, ..., R_{SB}\right) \\ \text{sous la contrainte d'égalité} \\ R_T\left(R_{LLL_D}, R_1, ..., R_i, ..., R_{SB}\right) = R_C \end{cases} \tag{2.37}$$

où R_T (i.e. le débit total) correspond à la somme pondérée des débits R_i et du débit de la basse fréquence R_{LLL_D}. SB est le nombre de sous-bandes produites par la TO3D

Notons que la sous bande-basse fréquence LLL_D est exclue du processus d'allocation. Cette sous-bande est codée par un simple codeur entropique afin de la transmettre avec une grande précision. En effet, cette sous-bande représente les moyennes locales de l'image, une erreur importante sur ces coefficients engendre une erreur importante sur de grandes zones de l'image reconstruite.

Ainsi en notant $R_i \mapsto D_i(R_i)$ la fonction distorsion-débit de la $i^{ème}$ sous-bande le problème de minimisation consiste maintenant à résoudre

$$\begin{cases} \hat{R}_T = \left(\widehat{R}_1, ..., \widehat{R}_{SB}\right) = \arg\min_{(R_1,...,R_{SB})} \sum_{i=1}^{SB} \beta_i D_i(R_i) \\ \text{sous la contrainte d'égalité} \\ \sum_{i=1}^{SB} \alpha_i R_i = R'_C = R_C - R_{LLL_D} \end{cases} \tag{2.38}$$

Avec α_i le rapport entre le nombre d'échantillons dans la sous-bande i et le nombre total de pixels de l'image. Les facteurs β_i dépendent de la non-orthogonalité des filtres et d'éventuelles pondérations psychovisuelles. $\hat{R}_T = \left(\widehat{R}_1, ..., \widehat{R}_{SB}\right)$ est le vecteur solution.

Ce problème d'optimisation peut être résolu en utilisant les opérateurs de Lagrange et en minimisant le critère J (appelé aussi fonctionnelle) :

$$J_\lambda(R_T) = \sum_{i=1}^{SB} \beta_i D_i(R_i) + \lambda \left[\left(\sum_{i=1}^{SB} \alpha_i R_i \right) - R'_C \right] \tag{2.39}$$

Une solution optimale au problème (2.38) peut être obtenue si nous savons déterminer un point-col de la fonction de Lagrange, c'est-à-dire un couple $(\hat{R}_T, \hat{\lambda})$ vérifiant les conditions de stationnarité données par les relations suivantes :

$$\begin{cases} J_{\hat{\lambda}}\left(\hat{R}_T\right) = \min_{R_\mathbf{T}} J_{\hat{\boldsymbol{\lambda}}}(R_T) \\ \text{avec} \\ \hat{\lambda}\left(\hat{R}_T - R'_C\right) = 0 \end{cases} \tag{2.40}$$

Soit S l'intervalle des solutions admissibles, nous pouvons définir la fonction $w(\lambda)$ pour $\lambda \geqslant 0$ par la relation (2.41). Cette fonction permet de connaître pour différents λ les débits $\hat{R}_i$ qui minimisent le critère $J_{\boldsymbol{\lambda}}(R_T)$.

$$w(\lambda) = \min_{R_\mathbf{T} \in S} J_{\boldsymbol{\lambda}}(R_T) \tag{2.41}$$

La recherche d'un point-col, s'il existe, peut se faire en résolvant le problème dual du problème initial (2.38) :

$$\begin{cases} \max_\lambda w(\lambda) = \max_\lambda \left\{ \min_{R_T \in S} J_\lambda(R_T) \right\} \\ \text{avec} \\ \lambda \in \mathbb{R} \end{cases} \tag{2.42}$$

Une solution permet de résoudre ce problème en deux étapes [86] :

- la première étape consiste à minimiser le critère Lagrangien par rapport à R pour un λ fixé. Nous disposons ainsi de l'allocation optimale des débits $\left\{\hat{R}_i \in \mathbb{R} \mid i = 1, ..., SB\right\}$. Cette étape garantit que l'on se situe sur l'enveloppe convexe de la fonction Débit-Distorsion $D(R)$, comme le montre la figure 2.17(a) ; Cette étape utilise une méthode basée sur le gradient conjugué [39].
- la seconde étape consiste à trouver le paramètre λ satisfaisant la contrainte d'égalité imposée sur le débit : $\hat{R}_T\left(\hat{R}_1, ..., \hat{R}_{SB}\right) = R_C$. Nous disposons ainsi du $\hat{\lambda}$ optimal parmi les solutions situées sur l'enveloppe convexe. Cette étape permet d'atteindre le débit cible R_C avec une allocation optimale des débits $\hat{R}_i$. La recherche du $\hat{\lambda}$ optimal est réalisée par une procédure itérative sur la fonction concave $w(\lambda)$ (voir figure 2.17(b)).

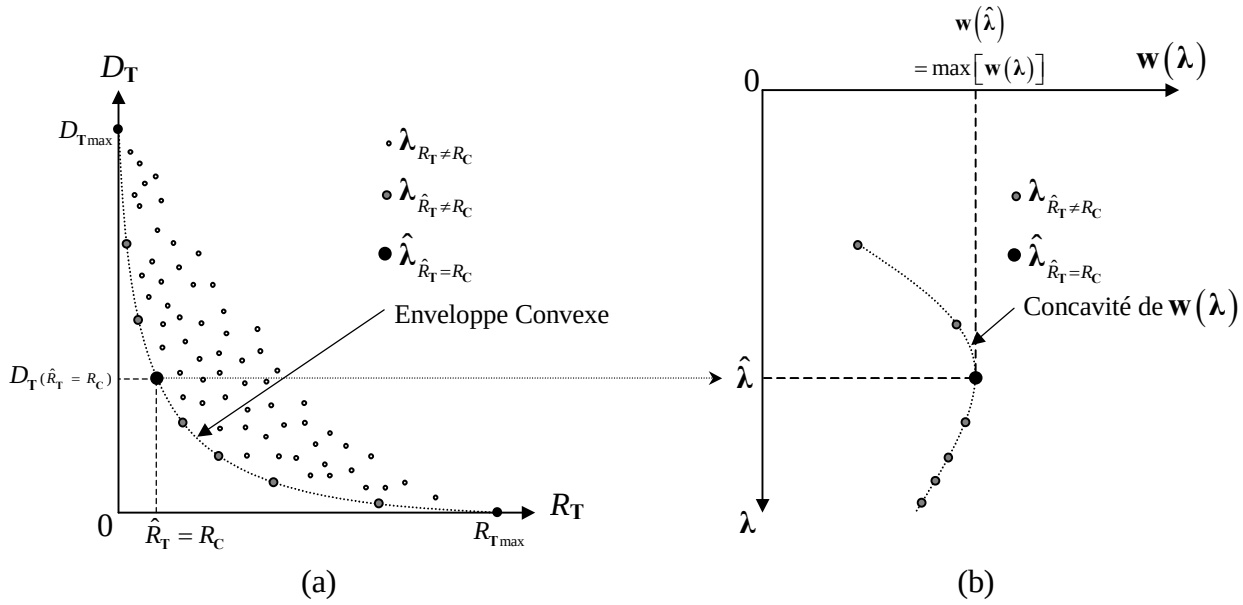

FIG. 2.17: Courbes $D_{\mathbf{T}}(R_{\mathbf{T}})$ et $\mathbf{w}(\boldsymbol{\lambda})$ pour l'allocation optimale des débits.

Dans ce paragraphe, nous avons proposé une méthode d'optimisation numérique de type lagrangienne pour effectuer l'allocation de débits des sous-bandes de la TO3D. L'augmentation de la complexité par l'ajout d'une variable supplémentaire (la zone morte) est compensée par l'utilisation d'un algorithme rapide quasi-optimal de recherche de la zone morte. Notons que ce schéma est différent de l'allocation proposée pour la QVAZM 2D dans [111]. Par ailleurs, si l'approche lagrangienne assure de trouver les paramètres de quantification optimaux assurant une très bonne précision, elle entraîne un temps de calcul qui peut s'avérer incompatible avec certaines applications médicales. Pour palier à cette contrainte, le chapitre 4 propose un simulateur limitant l'appel aux données utilisant une approximation des courbes débit-distorsion. Ce même chapitre propose une méthode encore plus rapide qui est quasiment sans appel aux données dans le processus d'optimisation. Elle s'inscrit dans la catégorie des méthodes par modèles statistiques.

2.6 Résultats expérimentaux

Nous avons testé la méthode QVAZM 3D présentée dans ce chapitre sur deux bases d'images médicales différentes. La première (E1) est la base décrite dans le tableau 1.1 du chapitre 1. Elle regroupe quatre scanners et quatre IRM de différents organes. Nous comparons notre méthode aux principales méthodes 3D de référence présentées dans le chapitre 1. La seconde base que l'on nommera E3 comprend quatre scanners ORL[5]. Les images compressées de cette base par différentes méthodes ont été évaluées par des médecins du Centre Alexis Vautrin de Nancy.

2.6.1 Résultats sur la base d'images E1

Nous avons tout d'abord testé notre algorithme sur la base d'images (E1) présentée dans le chapitre 1. Les images de cette base sont codées sur 8 bits alors que la tendance actuelle est à l'acquisition d'images sur 12 bits par les appareils d'imagerie médicale moderne (comme les scanners de la base E3). Néanmoins, comme nous l'avons déjà évoqué dans le chapitre d'introduction,

[5]Ces scanners ont été fournis par le Dr Philippe Henrot, Radiologue au Centre Alexis Vautrin de Nancy : Centre Régional de Lutte Contre le Cancer (C.R.L.C.C).

	CT1	CT2	CT3	CT4	MR1	MR2	MR3	MR4
QVAZM 3D	**34,69**	43,63	**41,08**	46,98	**38,84**	**36,61**	**39,03**	**42,98**
SPIHT 3D	33,98	**44,95**	40,51	**47,41**	38,37	36,31	38,76	42,97
QVA 3D	33,46	43,29	40,24	46,66	33,02	36,12	38,55	42,18

Tab. 2.1: PSNR moyen (dB) - Comparaison entre QVAZM 3D, SPIHT 3D et QVA 3D au débit de 0,1 bit/voxel sur les 8 images de la base E1.

	SPIHT 3D	QVAZM 3D	QVA 3D	CB-EZW 3D	ESCOT 3D	SPIHT 2D
0,5 bits/voxel	44,11	**44,14**	43,61	39,82	43,62	36,88
0,1 bits/voxel	33,98	**34,69**	33,46	31,68	34,68	26,34

Tab. 2.2: PSNR moyen (dB) - Comparaison entre les meilleurs algorithmes et notre méthode (QVZAM 3D) sur les 128 coupes de CT1 aux débits de 0,1 bit/voxel et 0,5 bit/voxel.

des travaux importants de la littérature [10], [19],[58], [118] ont utilisé cette base de données commune pour évaluer leurs performances de compression. La comparaison de la QVAZM 3D avec les méthodes actuelles en est grandement facilitée. Par ailleurs, il est admis que 8 bits de dynamique peuvent suffire pour interpréter une IRM. Enfin, les scanners de la base E1 représentent des structures osseuses qui peuvent là encore être analysés avec précision sur 256 niveaux de gris. Ainsi, les résultats sur cette base ont à la fois valeur de comparaison avec les méthodes de référence et restent exploitables malgré les 8 bits de dynamique.

Les tests de trois méthodes QVAZM 3D, QVA 3D et SPIHT 3D ont été réalisés sur la base E1 en faisant varier le débit de 0,05 à 1 bit/voxel. Les trois méthodes utilisent le filtre 9.7 avec une TO3D dyadique sur quatre niveaux ou sur trois niveaux quand la résolution le long de l'axe z est limitée ($T_z < 64$). La QVA 3D et la QVAZM utilisent le réseau $\mathbb{Z}^n$ auquel on a associé la norme L^1, et l'allocation de débits de type lagrangienne décrite dans le paragraphe 2.5.2. Les images codées avec SPIHT 3D [58] utilise une version classique de l'algorithme avec un codeur arithmétique basé sur contexte. Le tableau 2.1 montre qu'à bas débit, la QVAZM 3D est numériquement supérieure à la QVA 3D et à SPIHT 3D pour toutes les images de la base E1 à l'exception des scanners CT2 et CT4. Les résultats aux autres bas débits (jusqu'à 0,5 bit/voxel pour CT1 comme le montre la figure 2.18 et 0,3 bit/voxel pour les images restantes) conduisent à la supériorité de notre méthode. A plus haut débit, l'algorithme SPIHT 3D est légèrement meilleur que notre méthode en terme de PSNR car l'efficacité de la zone morte vectorielle décroît naturellement à de tels débits. Le tableau 2.2 résume les résultats produits par d'autres méthodes 3D sur CT1 qui est une image abondement utilisée dans la littérature pour l'évaluation de la compression avec pertes des images volumiques médicales. Dans ce tableau, les résultats de ESCOT 3D et CB-EZW 3D proviennent respectivement de [10] et [118]. Notre chaîne de compression présente des résultats là encore numériquement supérieurs à ces méthodes.

Nous avons évalué visuellement la qualité des images reconstruites sur cette base et constaté une amélioration de la qualité visuelle avec notre approche, en particulier à bas débit. Par exemple, si nous observons le zoom de la coupe 60 de CT1 pour un TC = 80 : 1 (0,1 bit/voxel) figure 2.19 : zoom des images 2.23, 2.24 et 2.25), nous pouvons constater que certains détails sont mieux préservés avec notre méthode que celle de SPIHT 3D. D'une façon générale, à de tels débits, les images sont moins floues avec notre méthode qu'avec SPIHT 3D. A moyen débit (par exemple à 0,5 bit/voxel), il semble plus difficile de constater des différences entre les deux méthodes. Notons tout de même qu'à ces débits la qualité globale des images reconstruites est très bonne. La fin du chapitre donne différentes coupes de la base E1 codées avec les deux méthodes, à bas et moyen débit.

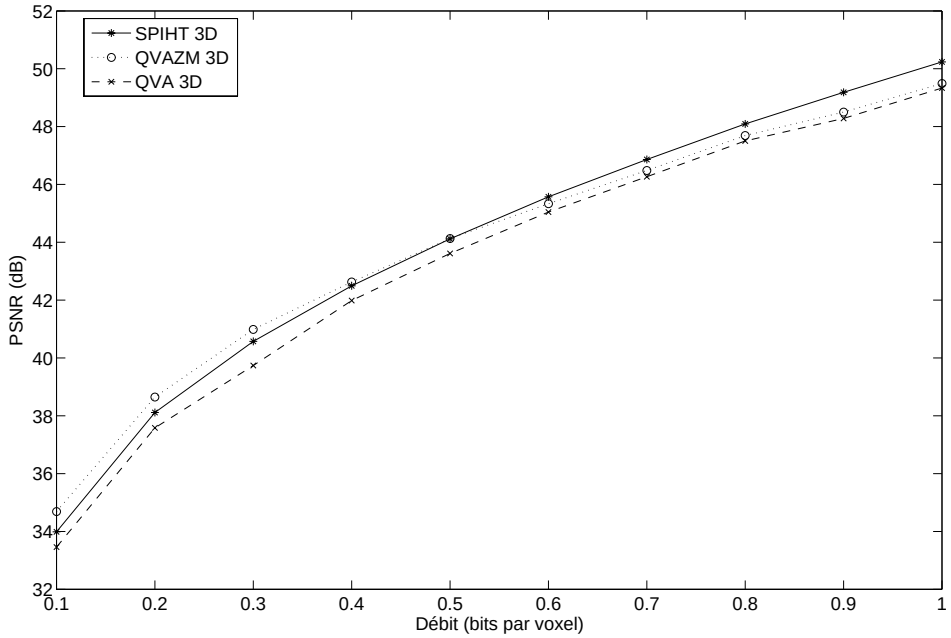

Fig. 2.18: Courbes PSNR moyen en fonction du débit de la QVAZM 3D, de la QVA 3D et SPIHT 3D - scanner CT1 (128 coupes).

2.6.2 Résultats sur la base d'images E3 (scanners ORL)

2.6.2.1 Conditions d'expériences

La seconde base (E3) testée comprend quatre scanners ORL codés cette fois-ci sur 12 bits au format DICOM. Ces images médicales 3D sont utilisées dans le diagnostic des pathologies ORL. Les caractéristiques de cette base sont décrites dans le tableau 2.3. Une étude a été mise en place avec deux spécialistes du Centre Alexis Vautrin de Nancy qui ont été chargés d'évaluer les effets de la compression sur la qualité des images décodées. L'originalité de l'étude provient du fait que l'équipe de médecins est composé d'un radiologue - Dr Philippe Henrot et d'un radiothérapeute - Dr Pierre Graff[6]. Le radiologue ayant une fonction de diagnostic demande des images reconstruites de très grande qualité. Le radiothérapeute quant à lui n'aura pas les mêmes exigences sur le niveau des images décodées. Notons enfin que la précision de diagnostic n'est pas évaluée dans cette étude car les quatre scanners de la base E3 correspondent à quatre patients sains.

La première étape de l'étude a consisté à compresser un scanner ORL de référence à différents taux de compression variant de 8 : 1 à 192 : 1 avec la méthode SPIHT 3D. Ce scanner de référence a permis de sélectionner une limite de compression à tester pour l'étude qui se situe pour ce type de modalité à environ 32 : 1. Au delà de cette limite, la qualité est jugée totalement inacceptable par les deux spécialistes. De plus, cette étape préliminaire a permis de sélectionner quatre coupes de références dans la pile d'images. Dans ces quatre coupes se trouvent des structures médicales importantes pour le praticien qui seront évaluées avec beaucoup d'attention. Ce scanner de référence n'est plus utilisé par la suite dans l'évaluation de la qualité. Les quatre coupes de références correspondantes pour chaque scanner de E3 ont été choisies à partir des critères définis à l'aide d'un médecin interne indépendant pour ne pas biaiser l'étude.

[6]L'étude a été initiée par le Dr Michel Lapeyre aujourd'hui en poste au Centre Jean Perrin de Clemont-Ferrand.

Modalité	Nom de la pile d'images	Partie du corps	Taille volume	Taille voxels (mm)
Scanner	A	Tête	$512 \times 512 \times 96$	$0,7 \times 0,7 \times 2$
Scanner	B	Tête	$512 \times 512 \times 96$	$0,17 \times 0,17 \times 2$
Scanner	C	Tête	$512 \times 512 \times 128$	$0,25 \times 0,25 \times 1$
Scanner	D	Tête	$512 \times 512 \times 128$	$0,35 \times 0,35 \times 2$

Tab. 2.3: Description des quatre piles d'images de l'étude E3.

Note	Signification
5	image de qualité excellente
4	image de très bonne qualité
3	image de qualité moyenne (médicalement acceptable)
2	image dégradée (médicalement inacceptable)
1	image très dégradée (totalement inacceptable)

Tab. 2.4: Description de la signification de chaque note de 1 à 5 pour l'évaluation de la qualité locale et globale des images de la base E3.

L'ensemble des images du tableau 2.3 a ainsi été compressé aux taux de 12 : 1 (1 bit/voxel), 16 : 1 (0,75 bit/voxel), 24 : 1 (0,5 bit/voxel) et 32 : 1 (0,375 bit/voxel) avec notre méthode QVAZM 3D ainsi que SPIHT 3D (avec des paramètres pour les deux méthodes identiques à ceux décrits dans le paragraphe 2.6.1). Les quatre coupes de références des quatre patients à tous les taux de compression sont anonymisées et totalement mélangées. Au total, nous avons 144 images 2D[7] à évaluer avec les deux méthodes de compression testées. Les deux spécialistes donnent une note de 1 à 5 pour ces 144 coupes 2D. Le tableau 2.4 résume la signification de chaque note pour l'appréciation de la qualité locale sur ces quatre coupes particulières. Par ailleurs, pour chaque coupe donnée, les neuf images 2D correspondant aux différents débits et méthodes testées sont mélangées pour éviter le phénomène de mémoire. Par la suite, les médecins donnent une note globale à la pile d'images complète compressée pour les différents taux de compression et méthodes testés. Dans cette note globale, les piles d'images sont également mélangées au hasard et la signification des notations est identique à celle de la qualité locale.

2.6.2.2 Résultats obtenus

D'un point de vue numérique, nous comparons les deux méthodes testées à travers le PSNR moyen en fonction du taux de compression choisi pour l'étude. Les résultats pour les quatre cas de l'étude sont résumés dans le tableau 2.5. Nous remarquons cette fois que l'algorithme SPIHT 3D est supérieur à notre méthode en terme de PSNR. Cependant, une notation par deux spécialistes qui utilisent quotidiennement ce type d'images nous apprendra beaucoup plus sur la qualité des images compressées.

Ainsi, le radiologue et le radiothérapeute ont évalué les images compressées suivant les conditions d'expériences décrites dans le paragraphe 2.6.2.1.

Les tableaux 2.6 et 2.7 présentent la moyenne des notations des quatre coupes de référence (notées C_1,C_2,C_3 et C_4) sur les quatre scanners de E3 en fonction de la méthode et du taux de compression, respectivement pour le radiologue et le radiothérapeute.

[7] 4 patients $\times$ 4 coupes de référence$\times$ (2 méthodes $\times$ 4 taux de compression +1 non compressée) = 144 coupes à évaluer.

	0.375 bits/voxel		0.5 bits/voxel		0.75 bits/voxel		1.0 bits/voxel	
Cas	(Q)	(S)	Q	S	Q	S	Q	S
Pile A	54.81	55.15	57.50	57.92	61.08	61.87	64.43	64.91
Pile B	55.42	55.61	57.74	58.24	61.59	64.85	64.51	64.85
Pile C	57.80	58.01	59.85	60.33	63.49	63.86	66.10	66.64
Pile D	59.00	59.05	61.44	61.54	65.10	65.35	67.84	68.40

Tab. 2.5: PSNR moyen (dB) - Comparaison entre SPIHT 3D (noté S) et la QVAZM 3D (notée Q) à différents débits testés sur les quatre scanners de la base E3.

Pour le radiologue, nous pouvons remarquer dans le tableau 2.6 que la moyenne générale des notes (dernière ligne du tableau) suit l'ordre des taux de compression indépendamment de la méthode utilisée. Les images originales sont clairement identifiées et jugées comme les images de meilleure qualité. Pour les taux de compression de 12 : 1 et 16 : 1, les images compressées restent de très bonne de qualité pour le radiologue. Le taux de compression 24 : 1 produit des images de qualité moyenne. Quant au taux de compression 32 : 1, il génère des images jugées très dégradées et ce, surtout pour la méthode SPIHT 3D. Si nous comparons, les deux méthodes de compression à travers l'évaluation du radiologue, la QVAZM 3D est en moyenne légèrement supérieure à SPIHT 3D sur l'évaluation locale des quatre coupes choisis à partir d'un taux de 16 : 1. De plus, cette supériorité visuelle devient assez significative pour le plus fort taux de compression testé (32 : 1). En ce qui concerne l'évaluation du radiothérapeute (tableau 2.7), la moyenne générale des notes suit de façon moins prononcée l'ordre des niveaux de compression. Par exemple, les images comprimées avec la QVAZM 3D à 12 : 1 sont jugées en moyenne de qualité équivalente aux images originales. De même, il n'y a pas de différence en moyenne entre le taux de compression 16 : 1 et 24 : 1 et cela quelque soit la méthode de compression. Par ailleurs, nous remarquons un jugement sur la qualité des images reconstruites (les piles) beaucoup moins sévère de la part du radiothérapeute que de celui du radiologue : les images sont jugées en moyenne de très bonne qualité jusqu'à un taux de 24 : 1 et encore médicalement acceptable à un taux de 32 : 1. Pour la comparaison des deux méthodes de codage, la QVAZM 3D apparaît encore une fois en moyenne légèrement supérieure à SPIHT 3D à tous les taux de compression. Par ailleurs, les figures 2.33 à 2.41 montrent différentes coupes de référence (C_i) compressées par les deux méthodes à différents taux de compression. Pour chaque image, nous précisons la note locale du radiologue et du radiothérapeute, respectivement notée n_{rad} et $n_{thp.}$

Les tableaux 2.8 et 2.9 fournissent les notations globales des deux spécialistes sur l'ensemble des coupes de chaque pile testée. Pour cette étude globale, le radiologue discrimine encore une fois bien les taux de compression suivant une logique assez similaire à celle de l'évaluation des quatre coupes de références. En effet, la limite de compression pour une très bonne qualité se situe à peu près à 16 : 1. Les deux autres taux de compression 24 : 1 et 32 : 1 produisent des images qui sont jugées respectivement dégradées et très dégradées. Par contre le radiothérapeute ne note pas de différences majeures entre les piles originales et celles compressées jusqu'à un taux de compression de 24 : 1. Par ailleurs, la comparaison entre les deux méthodes de compression donne encore une fois en moyenne un léger avantage à notre méthode. Cependant, pour la plupart des piles et des taux de compression, les deux méthodes présentent des résultats similaires. Enfin notons que la seule différence significative se situe sur la pile C à un taux de compression de 24 : 1 (note de 5/5 pour la QVAZM 3D, note de 1/5 pour SPIHT 3D).

Pour conclure, cette étude nous donne des premières informations sur l'appréciation de la compression avec pertes pour les scanners ORL. Les scanners ORL comprimées jusqu'à un taux de compression de 16 : 1 sont jugés localement et globalement de très bonne qualité et apparaissent compatibles avec la pratique clinique classique. Pour un taux de compression de 24 : 1, seul le

radiothérapeute juge les scanners de très bonne qualité pour sa pratique quotidienne. Par ailleurs, comme sur la base E1, la qualité visuelle issue de la méthode proposée dans ce chapitre apparaît légèrement supérieure à celle de SPIHT 3D (principalement pour l'évaluation locale des quatre coupes de référence). Enfin, soulignons que ce travail constitue une étude préliminaire qui va se prolonger en évaluant des cas supplémentaires afin de proposer une étude statistique détaillée (ressemblante à celle proposée dans le chapitre 3). De plus, nous avons vocation à utiliser d'autres critères d'évaluation. A ce sujet, une étude des effets de la compression sur la volumétrie (très utile en radiothérapie) dans les scanners ORL ainsi que sur des mesures anatomiques classiques pour cette modalité est en cours.

2.7 Conclusion et perspectives

Ce chapitre a proposé un nouveau schéma de compression avec pertes dédiés aux piles d'images médicales (scanner, IRM). la première étape du schéma s'appuie sur la transformée décoréllante reconnue comme la plus performante pour ce type de source : la transformée en ondelettes 3D (TO3D). La seconde étape correspondant à la quantification utilise une approche par blocs à travers une quantification vectorielle algébrique sur le réseau $\mathbb{Z}^n$ auquel on associe la norme L^1 et une zone morte vectorielle. Nous avons montré que pour ce type de quantificateur, une découpe des vecteurs avantageuse suit une orientation cubique $(2 \times 2 \times 2)$ pour capturer le phénomène d'agglutinement des coefficients d'ondelettes significatifs et inversement pour augmenter la concentration de vecteurs de faible norme. Ainsi, **un nouveau quantificateur** a été proposé **pour exploiter cette concentration de vecteurs de faible énergie**. Il constitue l'apport majeur de ce schéma de compression et s'appuie sur **une zone morte vectorielle** qui correspond à la cellule de Voronoï d'origine dilatée et modifiée permettant d'améliorer le compromis débit-distorsion. Pour les images médicales volumiques, le gain obtenu par le seuillage des grandes zones non significatives dans le domaine des ondelettes permet de **mieux coder les structures importantes des images**. Cependant, cette zone morte constitue un paramètre supplémentaire à régler dans le processus d'allocation de débits. Pour éviter d'augmenter la complexité, ce qui est capital en imagerie médicale, nous avons proposé un algorithme rapide de recherche de la zone morte quasi optimale qui se limite aux tests de quelques valeurs seulement. Ainsi, cet algorithme permet de limiter l'appel aux données au sein de la minimisation lagrangienne sous contrainte d'égalité qui est utilisée pour notre allocation de débits.

La méthode proposée : la QVAZM 3D donne des résultats numériques et visuels très encourageants. Sur des scanners et IRMs de différentes parties du corps (base E1), à bas débit, notre méthode s'avère supérieure numériquement aux principales méthodes de référence présentées dans le chapitre 1. Par ailleurs, elle présente une meilleure qualité visuelle que l'un des codeurs de référence pour ce type d'images : SPIHT 3D. De plus, une étude pour évaluer la compression avec pertes de notre méthode a été mise en place avec deux spécialistes du scanner ORL (un radiologue et un radiothérapeute). Les scanners ORL comprimés jusqu'à un taux de compression de 16 : 1 sont jugés localement et globalement de très bonne qualité et apparaissent compatibles avec la pratique clinique classique pour les deux médecins. Le radiothérapeute repousse lui cette limite d'acceptabilité de la compression jusqu'à un taux de compression de 24 : 1. Par ailleurs, comme sur la base E1, la qualité visuelle de la méthode proposée dans ce chapitre apparaît légèrement supérieure à SPIHT 3D.

Les perspectives de cette méthode sont nombreuses. Premièrement, la troisième étape de notre chaîne de compression correspondant au codage des index consiste en une estimation du débit entropique pour la norme et un débit binaire fixe pour la position. Proposer un codeur par plages qui possède une entropie inférieure à l'entropie ordre zéro peut permettre de réduire encore le débit. Deuxièmement, pour une méthode comme celle proposée ici, n'utilisant pas une approche par plans de bits successifs, les 12 bits de dynamiques peuvent s'avérer être un lourd handicap. En effet,

la TO 3D produit des coefficients d'ondelettes de plus grandes valeurs que sur des images codées sur 8 bits. Ainsi, certains vecteurs possèdent une norme très élevée entrainant un indexage de ces vecteurs sur des couches très peuplées qui peut s'avérer problématique en terme de complexité et de représentation des nombres. Proposer des solutions d'indexage pour ce verrou technologique est également une perspective intéressante. Troisièmement, le codage sans perte n'a pas été exploré dans ce manuscrit. Adapter notre méthode pour ajouter la fonctionnalité de compression sans perte est en cours. Cela consiste en une première partie identique à celle présentée dans ce chapitre suivie par un codage de l'erreur dans le domaine de l'image pour obtenir l'image originale. Enfin, nous avons vocation à expérimenter la QVAZM 3D sur d'autres modalités et à mettre en place de nouvelles procédures d'évaluations pour valider notre méthode.

TC	NC	12 : 1		16 : 1		24 : 1		32 : 1	
Coupes		Q	S	Q	S	Q	S	Q	S
C_1	5	3,75	3,75	3	2,75	3,5	2,75	2	1,25
C_2	5	4	4	4	3,5	2,5	2,5	2,25	1,25
C_3	5	4	3,75	4	3,75	3	2,75	2,25	1,75
C_4	5	3,75	4	3,5	3,75	2,75	3	2,5	1,75
Moyenne des 4 coupes	5	**3,88**	**3,88**	**3,63**	3,44	**2,94**	2,75	**2,25**	1,5

Tab. 2.6: Notation de la qualité locale des quatre coupes de références par **le radiologue**. Pour chaque coupe C_i, moyenne de la note sur les quatre patients de E3 en fonction du taux de compression et de la méthode - Méthode Q : QVAZM 3D - Méthode S : SPIHT 3D

TC	NC	12 : 1		16 : 1		24 : 1		32 : 1	
Coupes		Q	S	Q	S	Q	S	Q	S
C_1	4,75	4,75	4,75	4	3,75	4,25	4	4,25	2,75
C_2	4,5	4,5	4,5	4	4,25	4,75	4,25	3	3,25
C_3	4,5	4,75	4	4,5	4	4	4	2,75	2,75
C_4	4,75	4,5	4	4,5	4,5	4	4,25	3,5	3,5
Moyenne des 4 coupes	4,63	**4,63**	4,31	**4,25**	4,13	**4,25**	4,13	**3,38**	3,06

Tab. 2.7: Notation de la qualité locale des quatre coupes de références par **le radiothérapeute**. Pour chaque coupe C_i, moyenne de la note sur les quatre patients de E3 en fonction du taux de compression et de la méthode - Méthode Q : QVAZM 3D - Méthode S : SPIHT 3D

TC	NC	12 : 1		16 : 1		24 : 1		32 : 1	
Pile d'images (cas)		Q	S	Q	S	Q	S	Q	S
A	5	5	4	4	4	2	2	2	1
B	5	4	4	3	3	2	2	1	1
C	5	5	5	5	5	5	1	1	1
D	5	5	4	3	3	2	2	1	1
Moyenne des 4 piles	5	**4,75**	4,25	**3,75**	**3,75**	**2,75**	1,75	**1,25**	1

Tab. 2.8: Notation par **le radiologue** de la qualité globale des quatre piles d'images de E3 en fonction du taux de compression et de la méthode - Méthode Q : QVAZM 3D - Méthode S : SPIHT 3D

TC	NC	12 : 1		16 : 1		24 : 1		32 : 1	
Pile d'images (cas)		Q	S	Q	S	Q	S	Q	S
A	3	4	4	4	4	3	3	2	1
B	4	5	4	3	4	4	3	2	2
C	4	3	3	4	4	4	4	3	3
D	4	4	3	4	3	4	5	3	2
Moyenne des 4 piles	3,75	**4**	3,75	**3,75**	**3,75**	**3,75**	**3,75**	**2,5**	2

Tab. 2.9: Notation par **le radiothérapeute** de la qualité globale des quatre piles d'images de E3 en fonction du taux de compression et de la méthode - Méthode Q : QVAZM 3D - Méthode S : SPIHT 3D

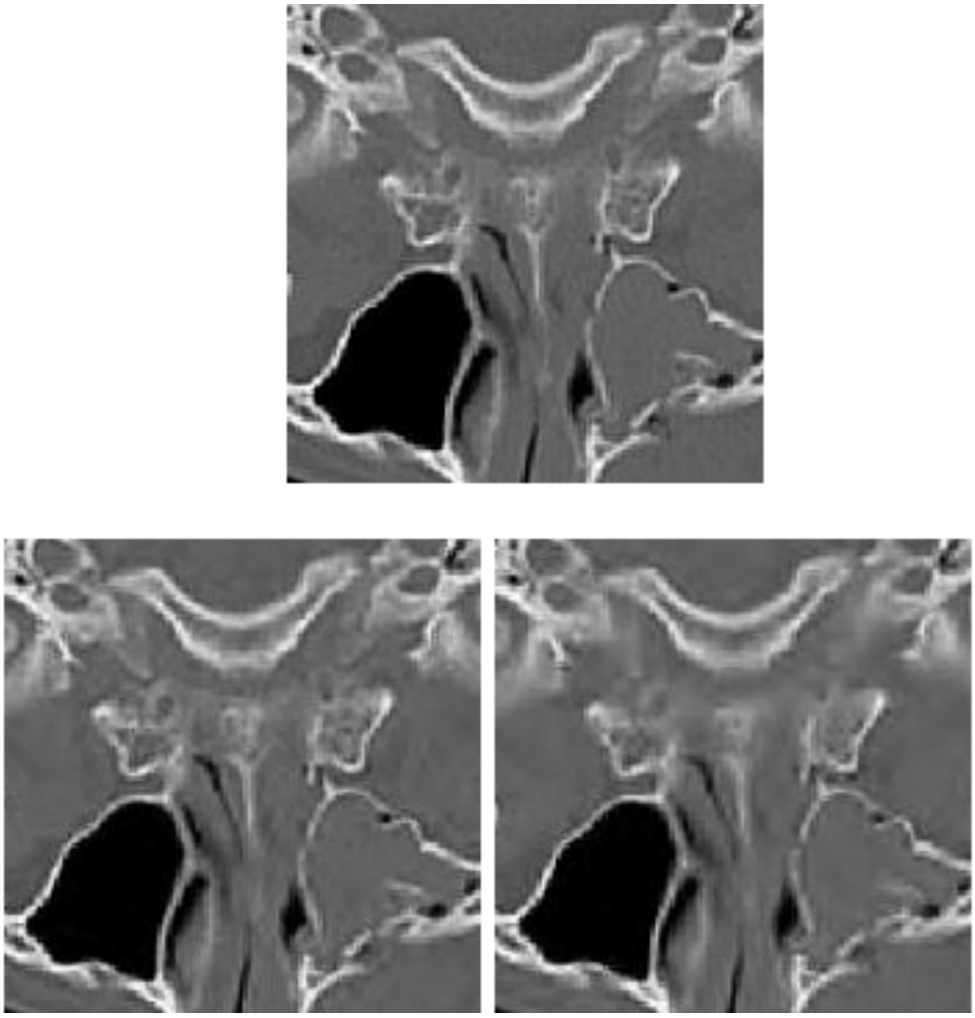

Fig. 2.19: Zoom de la coupe 60 décodée de CT1 - débit total de 0,1 bit/voxel pour les 128 coupes globales de CT1 - Dans le sens des aiguilles d'une montre en partant de l'image en haut : coupe originale, SPIHT 3D, QVAZM 3D.

Fig. 2.20: Coupe 30 originale de CT1.

Fig. 2.21: Coupe 30 de CT1 codée avec la méthode QVAZM 3D - débit total de 0,2 bit/voxel pour les 128 coupes.

Fig. 2.22: Coupe 30 de CT1 codée avec la méthode SPIHT 3D - débit total de 0,2 bit/voxel pour les 128 coupes.

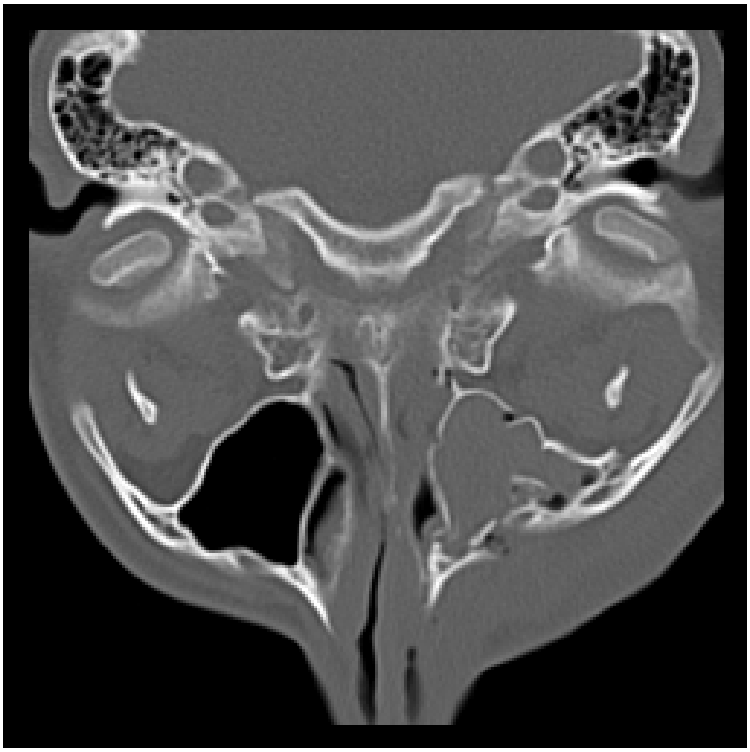

Fig. 2.23: Coupe 60 originale de CT1.

Fig. 2.24: Coupe 60 de CT1 codée avec la méthode QVAZM 3D - débit total de 0,1 bit/voxel pour les 128 coupes.

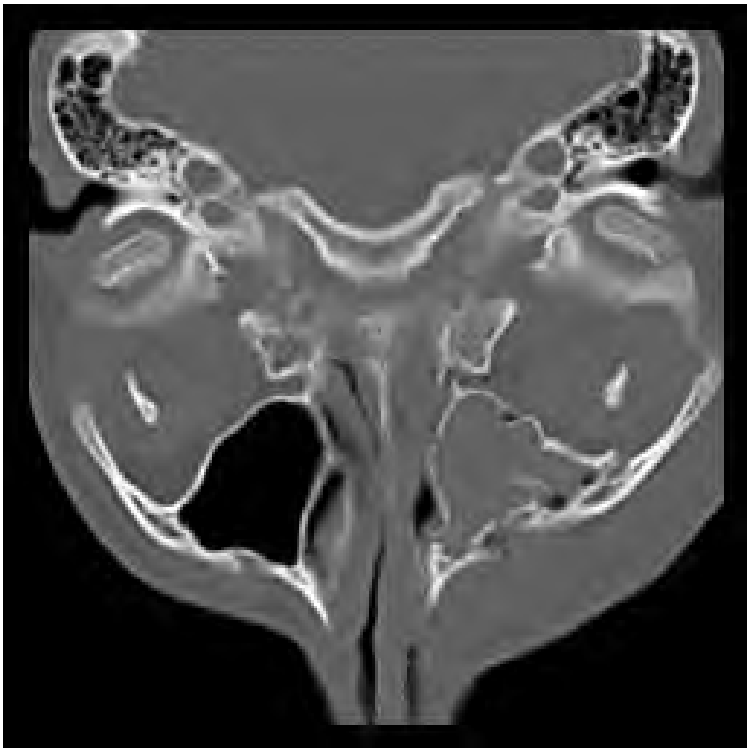

Fig. 2.25: Coupe 60 de CT1 codée avec la méthode SPIHT 3D - débit total de 0,1 bit/voxel pour les 128 coupes.

Fig. 2.26: Coupe 60 de CT1 codée avec la méthode QVAZM 3D - débit total de 0,5 bit/voxel pour les 128 coupes.

Fig. 2.27: Coupe 60 de CT1 codée avec la méthode SPIHT 3D - débit total de 0,5 bit/voxel pour les 128 coupes.

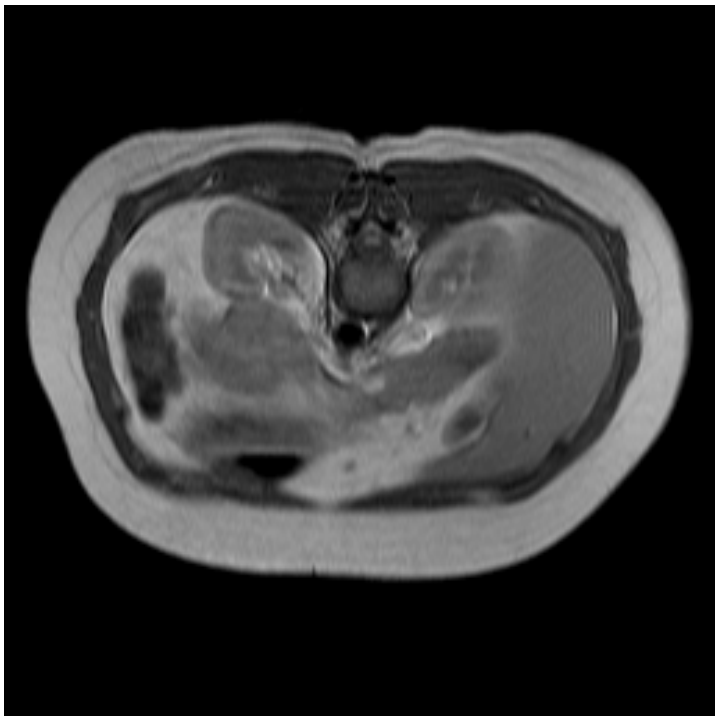

Fig. 2.28: Coupe 15 originale de MR2.

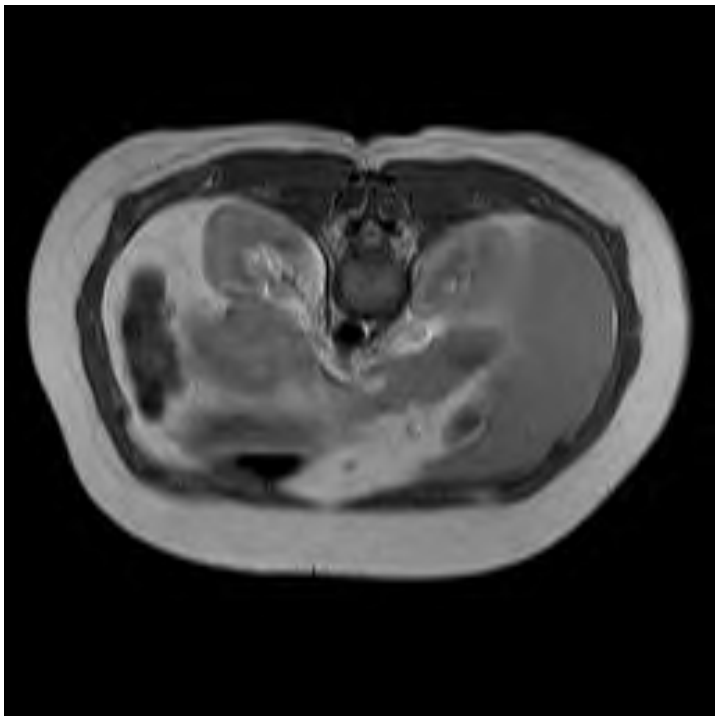

Fig. 2.29: Coupe 15 de MR2 codée avec la méthode QVAZM 3D - débit total de 0,1 bit/voxel pour les 48 coupes.

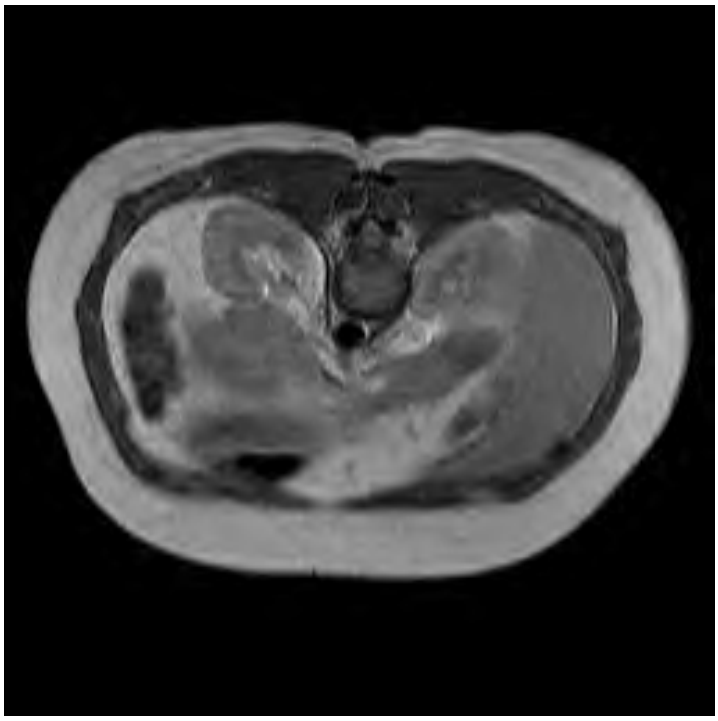

Fig. 2.30: Coupe 15 de MR2 codée avec la méthode SPIHT 3D - débit total de 0,1 bit/voxel pour les 48 coupes.

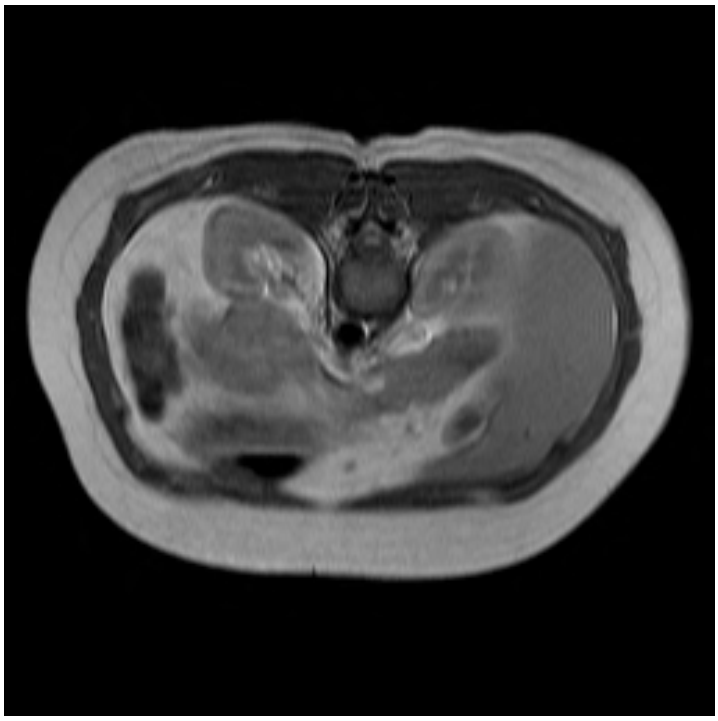

Fig. 2.31: Coupe 15 de MR2 codée avec la méthode QVAZM 3D - débit total de 0,5 bit/voxel pour les 48 coupes.

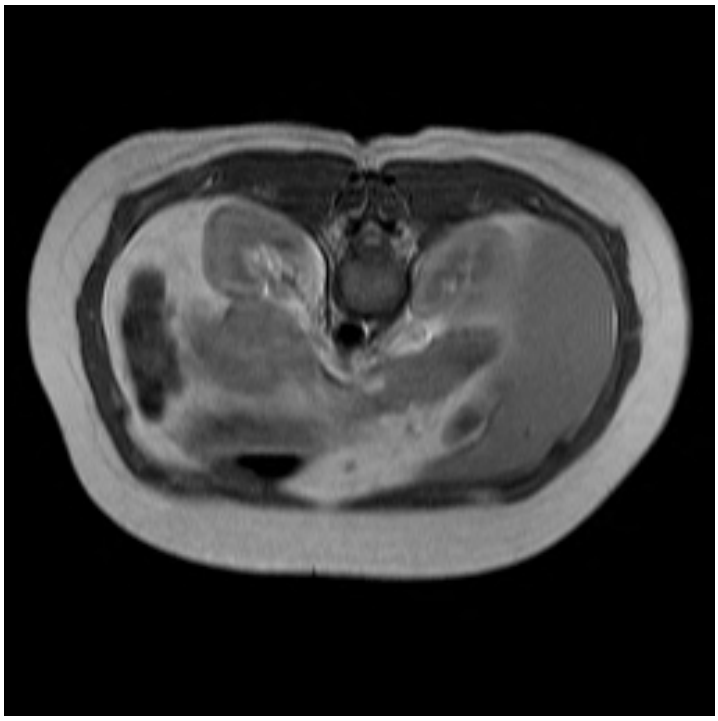

Fig. 2.32: Coupe 15 de MR2 codée avec la méthode SPIHT 3D - débit total de 0,5 bit/voxel pour les 48 coupes.

Fig. 2.33: Coupe C_1 originale de la pile d'images A. $n_{rad} = 5/5$ - $n_{thp} = 4/5$.

Fig. 2.34: Coupe C_1 de la pile d'images A codée avec la méthode QVAZM 3D - TC = 32 : 1 pour les 96 coupes de A. $n_{rad} = 1/5$ - $n_{thp} = 5/5$.

Fig. 2.35: Coupe C_1 de la pile d'images A codée avec la méthode SPIHT 3D - TC = 32 : 1 pour les 96 coupes de A. $n_{rad} = 1/5$ - $n_{thp} = 3/5$.

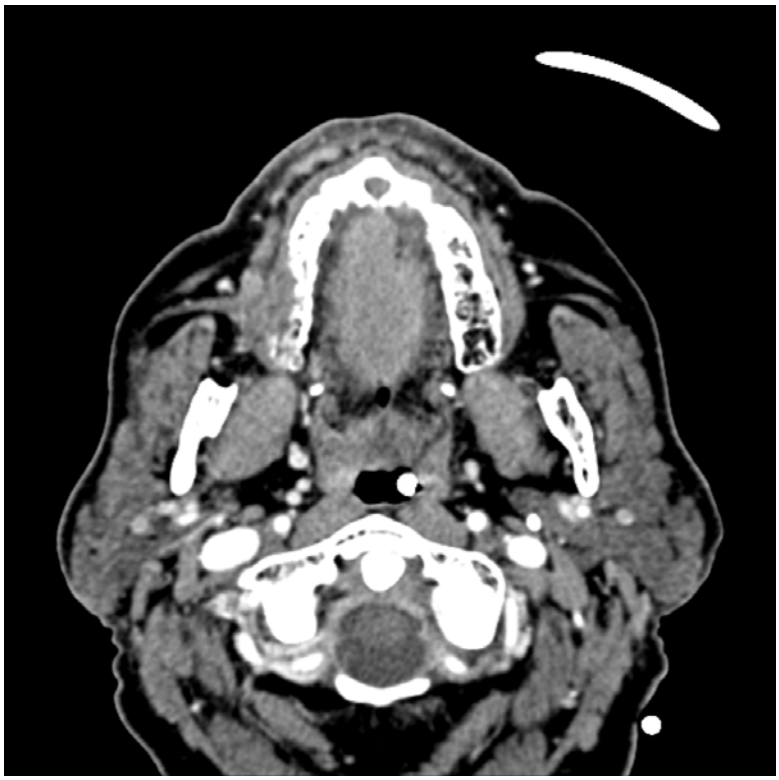

Fig. 2.36: Coupe C_2 originale de la pile d'images B. $n_{rad} = 5/5$ - $n_{thp} = 5/5$.

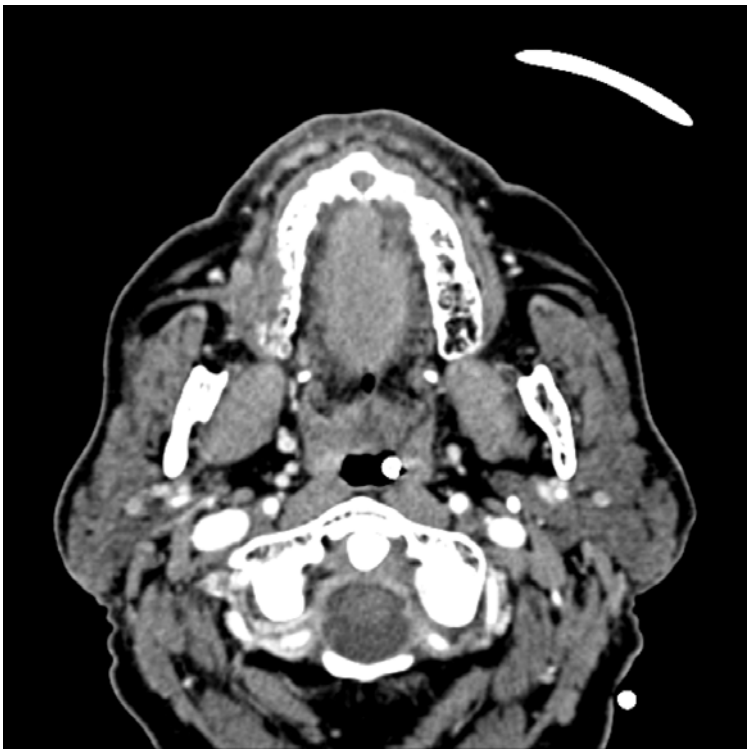

Fig. 2.37: Coupe C_2 de la pile d'images B codée avec la méthode QVAZM 3D - TC = 16 : 1 pour les 96 coupes de B. $n_{rad} = 4/5$ - $n_{thp} = 4/5$.

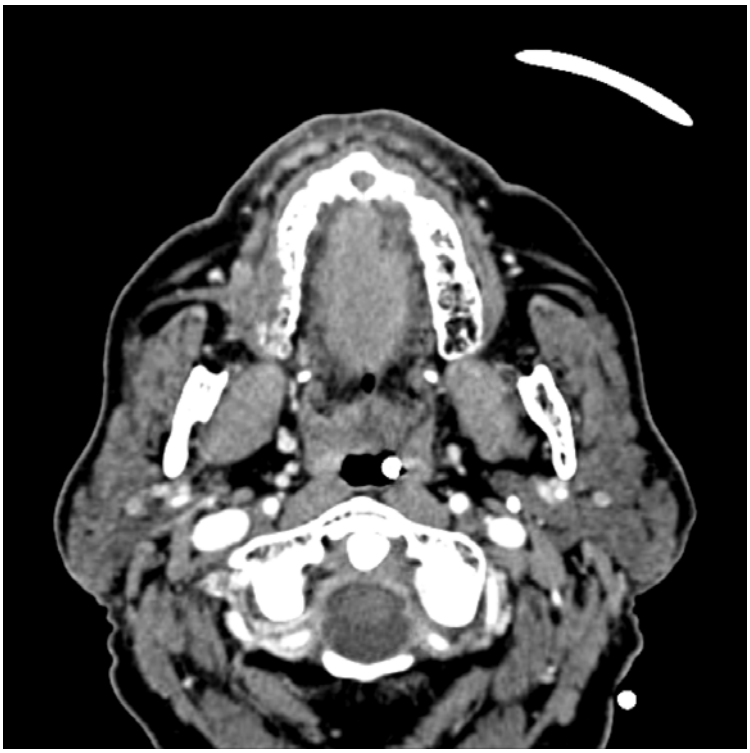

Fig. 2.38: Coupe C_2 de la pile d'images B codée avec la méthode SPIHT 3D - TC = 16 : 1 pour les 96 coupes de B. $n_{rad} = 3/5$ - $n_{thp} = 4/5$.

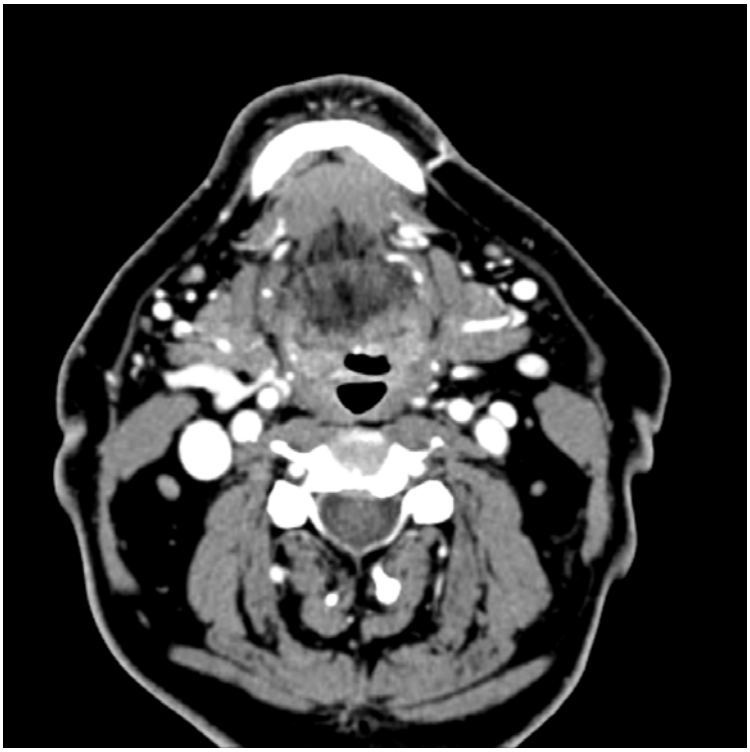

Fig. 2.39: Coupe C_4 originale de la pile d'images C. $n_{rad} = 5/5$ - $n_{thp} = 5/5$.

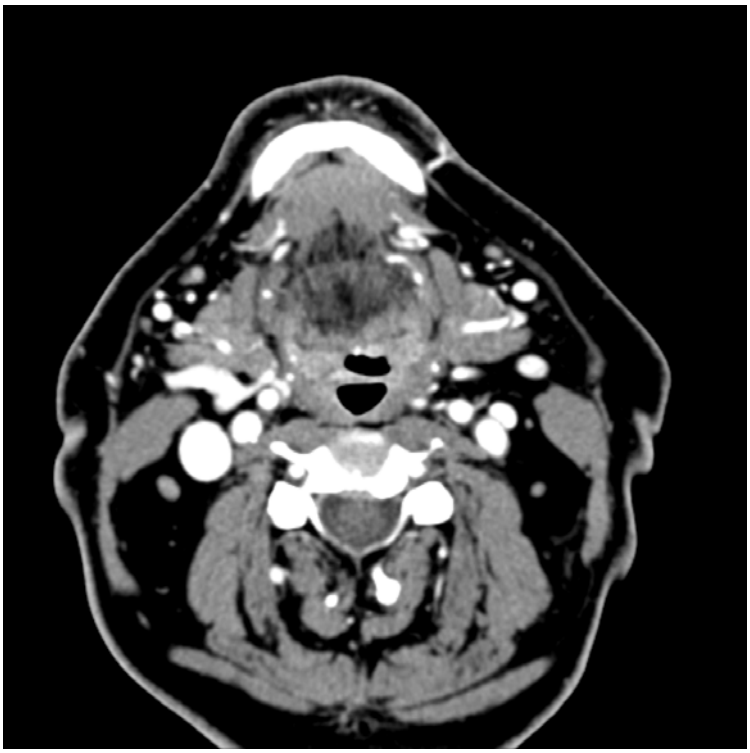

Fig. 2.40: Coupe C_4 de la pile d'images C codée avec la méthode QVAZM 3D - TC = 24 : 1 pour les 96 coupes de C. $n_{rad} = 4/5$ - $n_{thp} = 5/5$.

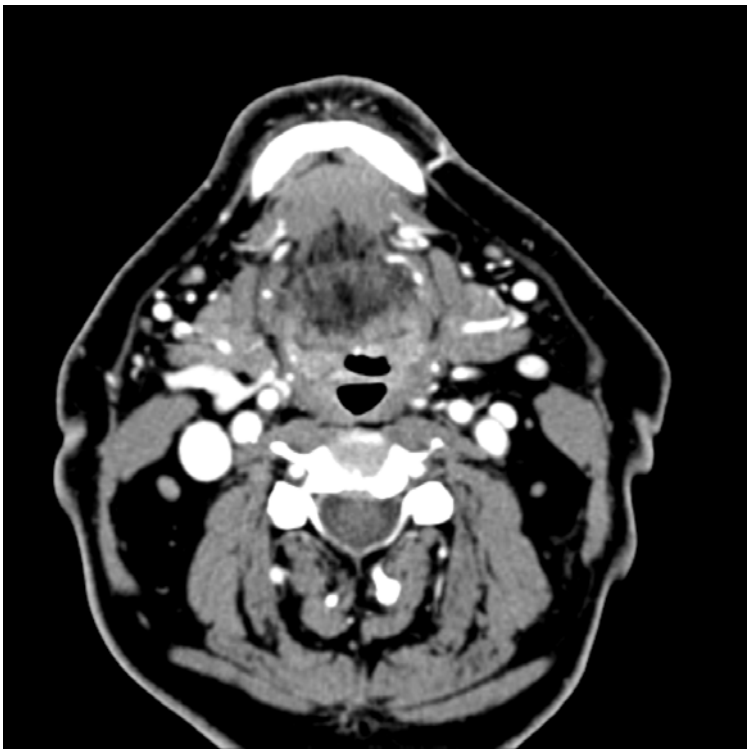

Fig. 2.41: Coupe C_4 de la pile d'images C codée avec la méthode SPIHT 3D - TC = 24 : 1 pour les 96 coupes de C. $n_{rad} = 4/5$ - $n_{thp} = 5/5$.

Chapitre 3

Evaluation des effets de la compression sur un outil d'aide à la décision pour les images médicales

3.1 Introduction

Le premier chapitre de ce mémoire a présenté plusieurs méthodes de compression avec pertes pour les images médicales 3D volumiques. Le suivant a proposé une méthode 3D originale pour la compression de ce type d'images. Les performances de ces méthodes ont été évaluées numériquement à travers le PSNR et visuellement à travers l'appréciation conjointe d'un radiologue et d'un radiothérapeute. Les résultats de cette étude originale d'évaluation nous a apporté une première indication sur la faisabilité de l'utilisation de la compression avec pertes pour les scanners ORL 3D à des taux de compression raisonnables. Par exemple, pour des taux de compression de 12 : 1, les images compressées sont jugées d'un niveau visuel équivalent voire supérieur aux images originales. Cela constitue un résultat pour un type de modalité lié à un organe particulier. Néanmoins, à notre connaissance, très peu d'études ont été réalisées sur les effets de la compression avec pertes pour les images médicales. Les seules études ([24], [44],[28]) assez anciennes que nous avons citées au chapitre 1 portaient sur des méthodes de compression à deux dimension. Ainsi l'évaluation des effets de la compression 3D sur les images médicales et sur les traitements d'images couramment utilisés dans le domaine médical demeure un problème ouvert.

Dans ce chapitre nous proposons d'évaluer les effets d'un algorithme de compression 3D [92], en l'occurrence SPIHT 3D [57], sur un outil d'aide à la décision ou CAD[1] [2]. Ce CAD permet la détection des nodules solides de poumons et les mesures de leur volumétrie. L'étude porte principalement sur **les effets de la compression sur la détection des nodules pulmonaires** [90], [92]. Quelques résultats sur la volumétrie [91] sont également présentés à la fin du chapitre. Le type de modalité lié à cette étude est le scanner de type MDCT[3]. Dans ce type de scanner, un tableau à deux dimensions d'éléments du détecteur remplace le vecteur d'éléments utilisé dans le scanner conventionnel. Ce tableau 2D permet au scanner d'acquérir des coupes multiples simultanément, de fournir des données haute résolution proche de l'isotropie et d'augmenter considérablement la vitesse d'acquisition. Dans le cadre des images MDCT de notre étude, les données haute résolution permettent d'augmenter la résolution de l'arbre broncho-vasculaire et d'afficher des nodules de poumons de très petite taille. Cependant, cette précision anatomique entraîne une expansion des

[1] Computer Aided Decision

[2] Ce travail a été réalisé en collaboration avec Philippe Raffy de la société R2Tech basée en Californie (http ://www.r2tech.com/main/home/index.php).

[3] MDCT : Multidetector computed tomography

volumes de données. Par ailleurs, le nombre de coupes fines à examiner et à stocker est constamment en augmentation (évaluation des nodules initiaux, examens complémentaires), accroissant le risque de rater la détection d'un nodule. Pour s'affranchir de ce risque, de nombreuses méthodes de CAD ont été développées pour la détection automatique des nodules pulmonaires dans les scanners. Ces systèmes ont également été créés afin de réduire la variabilité inter-observateur [4], [55]. Cependant, pour devenir des outils cliniques courants, les systèmes CAD doivent être intégrés dans les PACS des hôpitaux. Etant donné que ces systèmes requièrent un très grand nombre de coupes fines, une solution efficace est indispensable pour manipuler ces données 3D sans surcharger les réseaux de PACS. Dans ce contexte, on mesure l'intérêt d'évaluer la tolérance du CAD à la compression. L'influence des paramètres du scanner telle que la dose de radiation sur la performance de détection par des radiologues a déjà été étudiée [68]. D'autres études ont évalué la précision de la détection par un CAD pulmonaire en fonction de la dose de radiation et du filtre de reconstruction [3],[77]. Cependant, à notre connaissance, l'étude de performance de détection du CAD sur les données compressées avec pertes n'a pas été explorée.

Dans ce chapitre, nous commençons par présenter les matériels et méthodes utilisés pour l'étude. La deuxième partie est consacrée aux résultats suivie par les interprétations et les limitations. Nous terminons le chapitre en décrivant brièvement les premiers résultats sur les effets de la compression 3D sur la volumétrie. Enfin, notons que si ce chapitre est conjointement destiné à la communauté de la radiologie et à celle du traitement du signal, il comporte un grand nombre de termes médicaux dont les définitions ne sont pas intégralement détaillées. Nous nous en excusons par avance auprès des lecteurs non familiers au domaine médical.

3.2 Matériel et méthodes de l'étude

Ce paragraphe présente successivement la sélection des images qui constituent le standard doré de notre étude. Puis nous expliquons le choix de SPIHT 3D comme méthode de compression. Le principe et les fonctionnalités du CAD sont ensuite décrits brièvement. Enfin, nous présentons l'étude statistique mise en place pour tester l'influence de la compression sur la détection des nodules pulmonaires.

3.2.1 Sélection des images

Dans cette étude, les tests on été réalisés sur cent vingt MDCT anonymisés de poitrine d'adultes provenant de cinq sites différents des États-Unis. Nous décrivons l'ensemble des paramètres de ces images de l'étude car ils seront utiles dans l'étude statistique. La proportion des systèmes MDCT utilisés est la suivante : 30/120 sont obtenus à partir d'une solution Phillips Medical System, 34/20 utilisent GE Healthcare et 56/120 sont des scanners de type Siemens Medical Solutions. Les paramètres techniques du scanner consistent en un voltage du tube de 100 à 140 kV, une dose de radiation allant de 10 à 175 mA. La collimation qui peut être assimilée en simplifiant à l'épaisseur de coupe à l'acquisition varie de 1,0 à 1,25 mm. L'intervalle de reconstruction de l'image est compris entre 0,5 et 0,8 mm. Les algorithmes de reconstruction suivants sont considérés comme utilisant des noyaux de reconstruction haute-fréquence : "B70s", "B60f", "B60f" utilisés par Siemens, "E" de Phillips et "Lung" de GE. Les noyaux de reconstruction basse-fréquence sont composés de "B50f", "B30f" de Siemens, "L" de Phillips et "Standard" de GE.

Les cas de l'étude ont été sélectionnés pour avoir une large proportion de nodules entre 4 et 6 mm de diamètre comme le montre la figure 3.1. De plus, les images comportent un nombre limité de nodules par cas (en moyenne, deux par cas) et une proportion raisonnable entre les cas faiblement dosé et hautement dosé. La distribution des cas comprend une faible majorité de cas faiblement dosés (55%, 66/120), une large majorité de cas avec un noyau de reconstruction haute-fréquence

(91%, 109/120), et une minorité (22%, 26/120) de cas avec un média de contraste intraveineux. Le tableau 3.1 résume les fréquences et les pourcentages des variables étudiées.

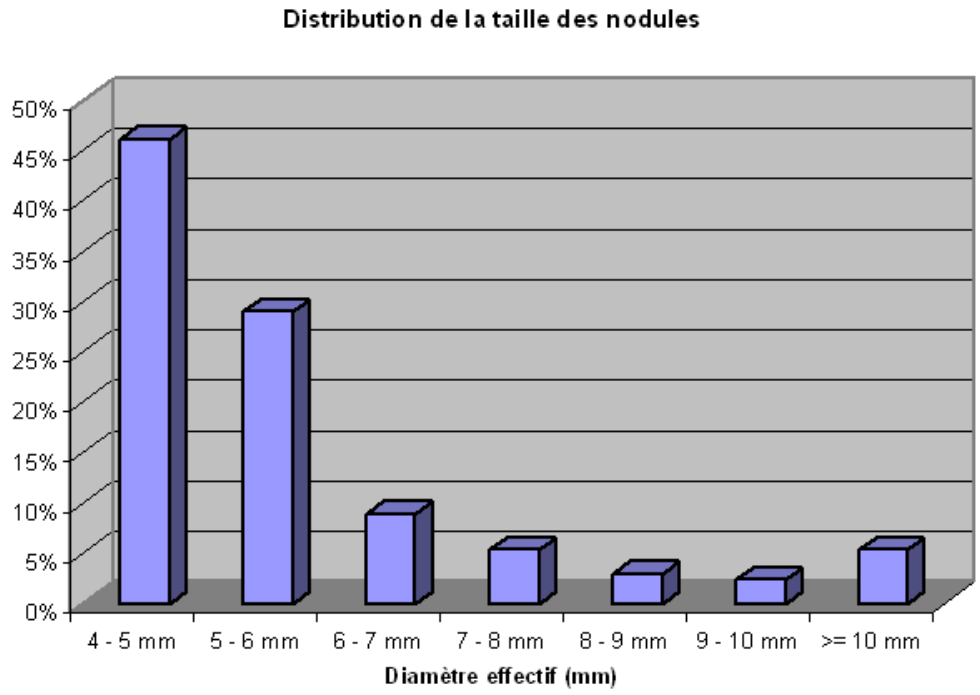

Fig. 3.1: Histogramme représentant la distribution des diamètres effectifs des nodules pour les 169 nodules de poumons de notre étude.

3.2.2 Standard doré

Pour mesurer la préservation de la sensibilité du CAD, il est nécessaire de déterminer un standard doré pour chaque cas qui servira de standard de comparaison avec les nodules détectés par le CAD pour chaque version compressée. Deux radiologues de l'université de Stanford expérimentés dans le domaine pulmonaire ont analysé indépendamment chaque cas non compressé pour les nodules de poumons, sans limite de temps d'interprétation. Nous avons demandé à chaque radiologue de marquer chaque nodule sur la plupart des coupes représentatives. Pour chaque nodule trouvé, les radiologues ont noté la potentielle activité, la densité et la calcification du nodule. Un nodule "potentiellement actif" est défini comme hautement suspicieux pour la malignité ou d'autres maladies, requiérant une intervention (opération, biopsie) ou un suivi d'au minimum un an. Ces nodules possèdent une densité moyenne plus grande que la densité entre la graisse et l'eau (entre -100 et 0 Hounsfield). Pour être diagnostiqué, des nodules pulmonaires potentiellement actifs doivent être définis comme un solide sphérique bien démarqué ou une ellipsoïde (longueur $\leq$ 3 fois largeur), ou ayant une opacité complexe et irrégulière. Les opacités linéaires, rectangulaires ou sous-pleurales sans apparence nodulaire sont exclues. C'est la cas par exemple d'un épaississement du mur bronchique, d'une consolidation du cubage focal et des surfaces affichant des artefacts de mouvement significatifs. Seuls les nodules avec un diamètre moyen plus grand que 4 mm et ayant au moins un quart de leur surface calcifié sont inclus dans l'étude. Les résultats de faible densité telles que les opacités "ground glass" ne sont pas pris en compte.

Variable	Comparaison	Pourcentage	
Dosage	faible ($\leq$ 60 mA)/ fort	faible : 55%(66/120)	fort : 45%(54/120)
Filtre de reconstruction	basse/haute fréquence	faible : 9,2%(11/120)	fort : 90,8%(109/120)
Media de constraste	oui / non	oui : 21,7%(26/120)	non : 78,3%(94/120)
Taille des nodules	petit ($\leq$ 5 mm) / large	petit : 46,2%(78/169)	large : 53,8%(91/169)
Localisation nodules	pleural / autres	pleural : 39,6%(67/169)	autres : 60,4%(102/169)

TAB. 3.1: Fréquences et pourcentages des variables étudiées

Le standard de référence pour chaque nodule potentiellement actif a été déterminé en incluant les résultats des deux radiologues. En cas de désaccord entre eux sur la présence ou la dangerosité potentielle d'un nodule détecté, la plus grande note (c'est à dire celle correspondant à un nodule potentiellement actif) a été gardée. Les cas sans nodule potentiellement actif définis par le standard de référence, ou avec des nodules plus petits que 4 mm, sont considérés comme négatifs (cas normaux).

3.2.3 Méthode de compression

Le but de l'étude est d'évaluer les performances du CAD sur des images compressées par un codeur d'images volumiques moderne. Nous avons utilisé l'algorithme SPIHT 3D [58] décrit dans le premier chapitre avec comme première partie une TO3D dyadique par lifting utilisant le filtre 9.7. Pour des raisons de complexité calculatoire, la décomposition 3D tout comme l'algorithme de codage est appliqué sur des groupes de seize coupes. En effet, les images 3D MDCT de notre étude comportent souvent plus de 500 images 2D. Tous les cas ont été compressés à des taux de 24 : 1, 48 : 1 et 96 : 1.

Le choix de l'algorithme SPIHT 3D est motivé par deux raisons principales. Premièrement, il constitue une référence pour la compression des images médicales 3D en terme de performance débit/distorsion. Deuxièmement, le seuillage des coefficients d'ondelettes non significatifs permet d'assigner plus de bits à ceux qui correspondent aux détails importants et aux fines structures des images de type MDCT. Enfin, au moment de cette étude, notre algorithme QVAZM 3D n'était pas finalisé. Quoiqu'il en soit l'objectif majeur de cette étude était plus de montrer la tolérance d'un CAD à la compression avec pertes que d'étudier les performances du CAD face à différentes méthodes de compression.

3.2.4 Système de CAD utilisé

Nous avons utilisé un système de CAD ayant une production commerciale : "ImageChecker CT CAD V 2.0" de la société R2 Technology Inc, Sunnyvale [12], CA. Ce CAD est conçu pour détecter les nodules solides plus grands que 4 mm sur les examens MDCT des poumons obtenus avec une collimation inférieure ou égale à 3 mm. Le système consiste en quatre modules principaux

- Un outil de segmentation des structures anatomiques principales tels que le mur thoracique, le médiastin, les artères et vaisseaux pulmonaires.
- Un outil de détection et d'analyse qui identifie les nodules candidats
- Un classificateur de type réseaux de neurones discriminant les nodules candidats sur des critères tels que la forme, la densité, la taille, la texture ...
- Une méthode de correction de volume appliquée aux images sélectionnées pour prévenir les effets disjoints de segmentation dans des structures voisines comme des vaisseaux, le mur thoracique et le médiastin, et des densités pulmonaires correspondant à des maladies (fibrose pulmonaire, emphysème[4], accumulation de fluide).

[4]L'emphysème est un gonflement ou une distension du tissu pulmonaire provoqué par la présence d'air piégé dans

Après avoir présenté les différents paramètres de l'étude, nous nous intéressons à l'étude statistique mise en place.

3.2.5 Analyse statistique

Une étude des effets de la compression sur la sensibilité du CAD pour la détection des nodules a été réalisée aux points de fonctionnement de 2,5 fausses marques (FM), 5 FM et 10 FM par cas en utilisant un test de MacNemar [21]. Les taux de fausses marques sont des substituts classiques à la spécificité dans les études médicales. Nous rappelons que la sensibilité et la spécificité sont définies dans le paragraphe des outils d'évaluations du chapitre 1. Une valeur de 2,5 FM/cas correspond au point de fonctionnement typique du système commercial dans cette étude. De plus, elle répond à la plupart des besoins pour l'interprétation quotidienne dans le diagnostic des examens de poumons. Nous avons arbitrairement étudié les effets de la compression à 5 FM/cas et 10 FM/cas pour deux raisons principales : la majorité des études sur des CAD dans la littérature ont reporté la sensibilité à des taux de fausses de marques plus élevés et nous reconnaissons le besoin dans des situations cliniques particulières (patients à haut-risque, cas compliqués) d'une sensibilité supérieure. Parallèlement, nous avons testé l'effet de la compression sur les taux de FM du CAD à des taux de sensibilité de 76,3%, 80,5% et 84,0% en utilisant un test de rang-signé de Wilcoxon.

Des modèles de régression logistique ont été utilisés pour évaluer l'impact du niveau de compression, du dosage (≤ 60 mA, > 60 mA), du filtre de reconstruction (basse fréquence, haute fréquence), du media de contraste intraveineux (≤ 5 mm, > 5 mm), et de la localisation des nodules (juxtapleuraux ou complètement entourés par le poumon aéré). Cette analyse est répétée à tous les points de fonctionnement (2,5 FM, 5 FM et 10 FM par cas). Pour rappel, la régression logistique se définit comme étant une technique permettant d'ajuster une surface de régression à des données lorsque la variable dépendante est dichotomique. Cette technique est utilisée pour des études ayant pour but de vérifier si des variables indépendantes peuvent prédire une variable dépendante dichotomique. Contrairement à la régression multiple et l'analyse discriminante, cette technique n'exige pas une distribution normale des prédicateurs ni l'homogénéité des variances.

Tous les tests ont été réalisés avec une hypothèse alternative bilatérale avec des P-valeurs inférieures à 0,05. Pour rappel, la P-valeur associée à un test est la probabilité d'observer la valeur obtenue si l'hypothèse nulle est vraie. Plus concrètement, elle correspond à une valeur statistique qui indique la probabilité qu'ont les observations contenues dans une étude à être dues au seul fait de la chance. Ainsi une P-valeur peut être utilisée comme repère de la confiance qu'on peut avoir dans un résultat particulier. Par exemple, une P-valeur égale à 0,05 signifie que le résultat observé dans l'étude peut se produire moins d'une fois en vingt études, par hasard.

Par ailleurs, en temps normal, lorsqu'on évalue les performances du CAD, les taux de FM du CAD sont déterminés sur les cas normaux. Pour notre étude, les taux de FM sont basés sur les cas jugés sans nodule potentiellement actif. Les conditions d'expériences de l'étude présentées, intéressons nous aux résultats produits par le CAD sur les 120 images compressées avec SPIHT 3D.

3.3 Résultats

La définition du standard doré a établi que 37 cas ne possédaient aucun nodule potentiellement actif. Les 83 cas restants contiennent 169 nodules potentiellement actifs, rangés dans une taille entre 3,8 mm et 35 mm en diamètre (moyenne 5,8 mm $\pm$ 3,0 (écart type) ; médiane 5,1 mm). Les nodules juxtapleuraux, définis comme des nodules touchant le mur thoracique, le médiastin ou le diaphragme avec une surface de contact d'au moins 10 %, représentent 40 % (67/169) de tous les nodules.

les interstices du tissu conjonctif ou du tissu inter-alvéolaire.

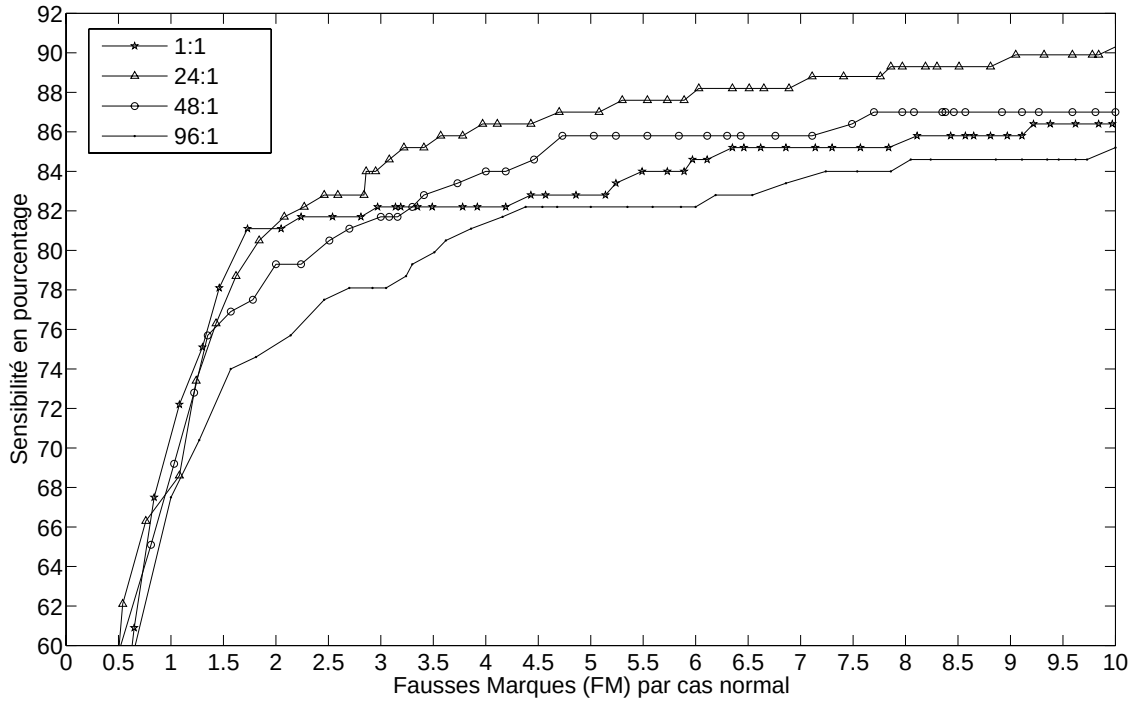

FIG. 3.2: Courbes FROC représentant la performance de détection du CAD des 169 nodules solides de poumons potentiellement actifs et plus grands que 4 mm à différents taux de compression.

Les courbes FROC[5] correspondant aux images non compressées, compressées aux taux de 24 : 1, 48 : 1 et 96 : 1 sont reportées sur la figure 3.2.

Les valeurs de sensibilité du CAD sont reportées dans le tableau 3.2. Si nous comparons les résultats obtenus avec les images originales non compressées, il n'y a pas de dégradation significative de la sensibilité à tous les points de fonctionnement et tous les taux de compression étudiés. Cependant, entre certains taux de compression, on constate une association marginale avec la sensibilité. Spécifiquement, les résultats pour un taux de compression de 24 : 1 sont meilleurs que ceux à des taux de 96 : 1 pour tous les points de fonctionnement, et meilleurs que ceux sans compression au point de fonctionnement de 10 FM/cas ($P = 0,03$).

Les taux de fausses marques par cas à différents niveaux de sensibilité de détection en fonction du taux de compression sont résumés dans le tableau 3.3. Les taux de FM pour tous les niveaux de

[5]La définition a été donnée au chapitre 1.

Taux de Compression	2,5 FM/cas	5 FM/cas	10 FM/cas
1 : 1	81,7 (138/169 ; 75,8 ; 87,5)	82,8 (140/169 ; 77,2 ; 88,5)	86,4 (146/169 ; 81,2 ; 91,6)
24 : 1	82,8 (140/169 ; 77,2 ; 88,5)	87,0 (147/169 ; 81,9 ; 92,1)	90,5 (153/169 ; 86,1 ; 94,9)
48 : 1	80,5 (136/169 ; 74,5 ; 86,4)	85,8 (145/169 ; 80,5 ; 91,1)	87,0 (147/169 ; 81,9 ; 92,1)
96 : 1	77,5 (131/169 ; 71,2 ; 83,8)	82,2 (139/169 ; 76,5 ; 88,0)	85,1 (144/169 ; 79,9 ; 90,6)

TAB. 3.2: Valeurs de la sensibilité du CAD en pourcentage à différents points de fonctionnement de fausses marques (FM) en fonction du taux de compression. Les nombres entre parenthèses sont le numérateur et le dénominateur suivis par les intervalle de confiance à 95 pct.

Taux de Compression	Sensibilité 76,3%	Sensibilité 80,5%	Sensibilité 84,0%
1 : 1	1,38 (51/37 ; 1,03 ; 1,81)	1,54 (57/37 ; 1,17 ; 2,00)	5,38 (199/37 ; 4,66 ; 6,18)
24 : 1	1,43 (53/37 ; 1,07 ; 1,87)	1,78 (66/37 ; 1,38 ; 2,27)	2,86 (106/37 ; 2,35 ; 3,47)
48 : 1	1,38 (51/37 ; 1,03 ; 1,81)	2,49 (92/37 ; 2,00 ; 3,05)	3,76 (139/37 ; 3,16 ; 4,44)
96 : 1	2,19 (81/37 ; 1,74 ; 2,72)	3,57 (132/37 ; 2,99 ; 4,23)	7,00 (259/37 ; 6,17 ; 7,91)

TAB. 3.3: Taux de fausse-marque (FM) du CAD à différents niveaux de sensibilité de détection. Les nombres entre parenthèses sont le numérateur et le dénominateur suivis par les intervalle de confiance à 95 pct.

compression tendent à augmenter quand on augmente la sensibilité. D'un point de vue statistique, les taux de FM à 96 : 1 sont significativement plus grands (moins bons) à tous les niveaux de sensibilité par rapport à ceux obtenus aux taux de compression inférieurs étudiés ($P < 0,01$). Les résultats des taux de FM sont moins bons sur les images comprimées à 48 : 1 que ceux sur les images originales à un niveau de sensibilité de 80,5% ($P = 0,03$), mais deviennent meilleurs pour un niveau de sensibilité de 84% ($P = 0,0002$). Similairement à l'analyse de la sensibilité, un taux de compression 24 : 1 fournit des taux de FM significativement inférieurs à ceux de 96 : 1 à toutes les sensibilités, et est meilleur que sans compression à un niveau de sensibilité de 84% ($P < 0,01$).

Les performances de détection du CAD commencent un peu à se dégrader à partir d'un taux de compression de 48 : 1. Bien que non étudiée, notre impression est qu'à mesure que la distorsion introduite par la compression augmente, la qualité d'image reconstruite baisse plus rapidement que la sensibilité. Ainsi, bien que la dégradation pour les performances du CAD ne soit pas statistiquement significative pour un taux de compression de 96 :1, la qualité visuelle pour ce niveau de compression est clairement en deçà de l'acceptation clinique. Le flou extrême des images combiné à la discontinuité d'un grand nombre de structures vasculaires est probablement responsable de l'augmentation des fausses marques du CAD. La proportion de fausses marques à l'endroit des vaisseaux normaux a augmenté de plus du double, alors que l'analyse statistique n'a indiqué aucun effet significatif ($P = 0,16$; test signé de rang de Wilcoxon). La proportion de toute autre cause de fausses marques, par exemple des cicatrices, des fissures, des anomalies pleurales, et des vaisseaux pulmonaires demeure inchangée.

Le tableau 3.4 présente un modèle de régression logistique pour évaluer les effets des variables présentées dans le tableau 3.1 sur le taux de détection des nodules à 2,5 FM/cas. Il présente le rapport de chance des différentes variables testées, suivi par l'intervalle de confiance à 95 % et la P-valeur. Le rapport de chance est la probabilité de vrais-positifs trouvés si la condition est présente divisée par la probabilité de vrais-positifs si la condition n'est pas présente. Par exemple dans ce même tableau, les nodules pleuraux ont en moyenne environ deux fois moins de chance d'être détectés par le CAD que d'autres types de nodules. A différents taux de compression, les tableaux 3.5 et 3.6 décrivent la sensibilité à 2,5 FM/cas respectivement en fonction des caractéristiques des nodules et des spécificités des cas. En se basant sur ces tableaux obtenus par l'analyse multivariables à 2.5 FM/cas la localisation du nodule est un prédicteur significatif ($P = 0,01$). Les nodules ayant un localisation juxtapleurale ont une sensibilité inférieure aux autres nodules. La signification de la localisation du nodule tend à disparaître à partir de taux de fausses-marques plus élevés ($P = 0,03$ à 5 FM/cas, $P = 0,13$ à 10 FM/cas) et à des taux de compression plus bas ($P = 0,005$ à 96 : 1, $P = 0,06$ à 48 : 1, $P = 0,04$ à 24 : 1, $P = 0,22$ à 1 : 1)

Les autres variables testées, à savoir la taille des nodules (petite, ≤ 5 mm), le dosage (faible, ≤ 60 mA), le filtre de reconstruction (basse fréquence) et le media de contraste intraveineux ne sont pas des prédicateurs significatifs quelque soit le taux de compression ou le point de fonctionnement.

Variable	Rapport de chance	95 % IC	P valeur
24 : 1 vs 1 : 1	1,09	0,82-1,45	0,56
48 : 1 vs 1 : 1	0,92	0,67-1,27	0,62
96 : 1 vs 1 : 1	0,77	0,55-1,06	0,11
Dose $\leq$ 60 mA	1,21	0,58-2,55	0,61
Filtre de reconstruction	0,83	0,35-1,97	0,67
media de contraste intraveineux	0,74	0,27-2,55	0,56
Taille du nodule $\leq$ 5 mm	0,58	0,27-1,22	0,15
Nodule pleural	0,41	0,20-0,83	0,01

TAB. 3.4: Modèle de régression logistique : effet des variables sur le taux de détection des nodules (2,5 FM/cas).

	A : Taille des nodules		
Taux de Compression	$\leq$ 5 mm en diamètre	$>$ 5 mm diamètre	Sensibilité différence (%)
1 : 1	79,5 (62/78)	83,5 (76/91)	-4,0
24 : 1	80,8 (63/78)	84,6 (77/91)	-3,8
48 : 1	78,2 (61/78)	82,4 (75/91)	-4,2
96 : 1	75,6 (59/78	79,1 (72/91)	3,5
	B : localisation des nodules		
Taux de Compression	Pleural	¬ Pleural	Sensibilité différence (%)
1 : 1	77,6 (52/67)	84,3 (86/102)	-6,7
24 : 1	76,1 (51/67)	87,3 (89/102)	-11,2
48 : 1	74,6 (50/67)	84,3 (86/102)	-9,7
96 : 1	67,2 (45/67)	84,3 (86/102)	-17,1

TAB. 3.5: Sensibilité (en pourcentage) en fonction des caractéristiques des nodules (taille, localisation) à 2,5 FM/cas. Les données entre parenthèses sont utilisées pour calculer la valeur de la sensibilité

	A : Dosage		
Taux de Compression	$\leq$ 60 mA	$>$ 60 mA	Sensibilité différence (%)
1 : 1	80,0 (56/70)	82,8 (82/99)	-2,8
24 : 1	84,3 (59/70)	81,8 (81/99)	+2,5
48 : 1	84,3 (59/70	77,8 (77/99)	+6,5
96 : 1	81,4 (57/70)	74,8 (74/99)	+6,6
	B : filtre de reconstruction		
Taux de Compression	Basse freq.	Haute freq.	Sensibilité différence (%)
1 : 1	75,8 (25/33)	83,1 (113/136)	-7,3
24 : 1	78,8 (26/33)	83,8 (114/136)	-5,0
48 : 1	78,8 (26/33)	80,9 (110/136)	-2,1
96 : 1	75,8 (25/33)	77,9 (106/136)	-2,1
	C : Media de contraste		
Taux de Compression	Contraste	$\neg$ Contraste	Sensibilité différence (%)
1 : 1	80,8 (21/26)	81,8 (117/143)	-1,0
24 : 1	80,8 (21/26)	83,2 (119/143)	-2,4
48 : 1	69,2 (18/26)	82,5 (118/143)	-13,3
96 : 1	73,1 (19/26)	78,3 (112/143)	-5,2

Tab. 3.6: Sensibilité (en pourcentage) en fonction des caractéristiques des patients (dosage, filtre de reconstruction, media de contraste) à 2,5 FM/cas. Les données entre parenthèses sont utilisées pour calculer la valeur de la sensibilité.

L'ensemble des résultats produits par l'analyse statistique amène à un certain nombres de discussions présentées dans le paragraphe suivant.

3.4 Discussion

Dans notre étude, qui a impliqué 120 examens MDCT de poitrine, nous avons montré que la compression avec pertes par SPIHT 3D ne détériorait pas la précision de détection du CAD jusqu'à un taux de compression de 48 :1. Nous supposons que cela est principalement dû à l'effet de débruitage du codeur 3D testé et à sa capacité à préserver les structures importantes dans l'image. Utiliser un codeur d'ondelettes très performant comme SPIHT est très bénéfique en terme de débruitage. D'un point de vue théorique, le schéma de compression qui a le meilleur compromis débit-distorsion possède aussi le meilleur modèle et le meilleur estimateur. Bien que théoriquement possible, un tel schéma n'est pas réalisable pratiquement puisqu'il utilise des tailles de blocs de longueur infinie. Cependant, plus le schéma de compression fonctionne près de la borne de compression indiquée par la fonction débit-distorsion de Shannon, meilleur sera le modèle qu'il incarne et meilleures seront les estimations qu'il produit [62], [88]. A mesure que le taux de compression augmente, le premier changement perceptible est typiquement la suppression du bruit "sel et poivre". Cet effet est particulièrement apparent pour des cas à faible dosage et est responsable de l'amélioration de la détection du CAD pour deux raisons principales. En premier lieu, le débruitage réduit la classe des fausses marques du CAD interprétées comme des nodules correspondant à l'agglutinement de voxels bruités, spécialement dans le parenchyme du poumon inférieur à coté du diaphragme. En second lieu, il augmente la sensibilité en enlevant des jonctions artificielles entre un nodule et ses structures voisines, telles que la plèvre ou certains vaisseaux (figure 3.3).

En dehors d'un CAD, des études ont été menées sur la tolérance des MDCT de poitrine à la compression. Ces évaluations utilisent des méthodes de compression 2D comme la quantification vectorielle [24], JPEG [60] et JPEG2000 [56]. Elles se concentrent principalement sur l'effet de la compression sur la sensibilité de détection des radiologues. En outre, il y a des différences majeures dans la conception de ces études. Les études [24] et [60] utilisent des coupes de haute-résolution spatiale (comme c'est le cas dans notre étude) alors que l'étude [56] concerne un balayage complet du scanner de poitrine. Ces études précédentes ont montré que l'exactitude du diagnostic pour la détection radiologique des nodules de poumons était préservée jusqu'à des taux de compression avec pertes de 10 : 1 [24], [54], [56], [60]. Cependant ces résultats sont obtenus à partir de méthodes 2D, il est donc probable que la capacité des algorithmes de compression 3D permettant d'exploiter les corrélations à travers les coupes de la pile des images axiales repousse cette limite plus loin. La figure 3.4 montre la moyenne, sur 30 MDCT représentatifs, des PSNRs moyens des coupes 2D d'un cas à différents taux de compression. Elle confirme encore une fois le bénéfice d'utiliser une méthode 3D, spécialement pour les forts taux de compression. Au delà de 24 : 1, SPIHT-3D permet quasiment de doubler le taux de compression comparé à SPIHT 2D pour une qualité globale équivalente en terme de PSNR. Nous rappelons que les calculs de distorsion comme le PSNR ne reflètent pas toujours la qualité perceptuelle.

Dans notre approche, à la différence des études précédentes utilisant des observateurs humains [56] [97], nous avons trouvé une sensibilité inférieure au CAD pour des nodules juxtaposant des surfaces pleurales particulièrement à fort taux de compression ou à des faibles taux de FM (tableau 3.4). Ceci démontre une légère altération de la segmentation des nodules et des poumons avec la compression, affectant la plupart du temps la détection de petits nodules ($\leq$ 5 mm). Par ailleurs, en se basant sur l'analyse multi-variable, le dosage ne s'avère un facteur prédictif significatif. Comme suggéré dans le tableau 3.4 et confirmé avec les régressions logistiques à différents taux de compression, nous notons une petite tendance à une augmentation de la sensibilité avec la compression pour des cas faiblement dosés, et cela même si le dosage n'est pas un prédicteur significatif. Ces

résultats suggèrent que dans le cas de la compression avec pertes des données MDCT, la dose de rayonnement pourrait être réduite (< 60 mA) sans compromettre la détection du CAD.

Dans la pratique, les images à faible résolution sont souvent archivées dans le PACS pour compenser le large volume de données acquis avec des examens de type MDCT. Cependant, il est préférable de fournir aux algorithmes du CAD les données en haute-résolution pour que la détection soit plus précise. Notre étude suggère qu'une méthode de compression 3D par ondelettes avec pertes peut être exécutée sans baisse significative de la détection du CAD, tout en minimisant considérablement le volume de données.

La compression d'image fournit un mécanisme pour utiliser plus efficacement l'espace mémoire d'archivage et la bande passante du réseau. Que le CAD soit utilisé à l'intérieur d'un PACS ou employé comme un système autonome à l'intérieur ou bien en dehors des réseaux d'hôpitaux, la tolérance du CAD à la compression avec pertes est bénéfique pour deux raisons principales. D'abord, en augmentant le stockage disponible, la compression avec pertes permet de réduire le coût d'exécution (coût de stockage) facilitant ainsi l'intégration du CAD dans un PACS, ou d'augmenter la capacité de stocker un plus grand historique des examens, améliorant de ce fait la comparaison avec les examens antérieurs. En second lieu, les données peuvent être ainsi stockées sous forme comprimée, envoyées sous ce format, et décodées au poste de travail de réception, réduisant de ce fait les demandes en bande-passante et économisant du temps de transmission. Cela est particulièrement utile dans une application de téléradiologie où la capacité de canal est limitée.

Par ailleurs, notre étude a plusieurs limitations. D'abord, nous avons restreint la définition des nodules aux nodules solides non calcifiés pour lesquels le logiciel de CAD des poumons a été conçu. Ces résultats devraient être confirmés pour d'autres applications liées aux poumons. En second lieu, les cas de l'étude ont été acquis avec une collimation de 1,0 - 1,25 mm, qui semble être un paramètre classique d'acquisition actuellement. Cependant, il est fort probable que la corrélation augmente avec l'utilisation de coupes plus fines (ex. 0,5 mm ou 0,75 mm), rendant l'utilisation d'un algorithme de compression 3D avec une TO3D bien plus efficace. Réciproquement, une acquisition plus grossière des coupes devrait tendre à avoir l'effet opposé. Troisièmement, la base de données utilisée dans cette étude ($n = 120$) est composée d'un mélange de cas extraits à partir d'une plus grande base de données d'apprentissage pour former l'algorithme du CAD ($n = 85$), et de cas aléatoires pour examiner son exécution ($n = 35$). Dans le meilleur des cas, la base de données devrait seulement être composée des cas de tests pour éviter une évaluation trop optimiste de l'exécution de l'algorithme du CAD. Nous croyons, cependant, que les résultats produits dans cette étude sont conservatifs. En effet, l'incorporation des cas d'apprentissage dans la base de données globale a plutôt tendance à produire un biais favorable pour les ensembles de données non comprimés.

Pour conclure, les performances de détection des nodules solides de poumons plus grands que 4 millimètres en taille n'ont pas souffert de la compression par SPIHT 3D jusqu'à 48 : 1. De plus, le CAD s'est avéré robuste jusqu'à un taux de compression de 96 : 1, et cela alors même que l'aspect visuel des images comprimées était fortement dégradé. Ces résultats suggèrent que la compression pourrait faciliter l'implantation d'un CAD dans un environnement de type PACS. Bien que ces résultats soient encourageants, l'impact du CAD comme deuxième radiologue employant des données comprimées reste à évaluer dans une étude avec des radiologues ou dans une épreuve clinique.

Le CAD pulmonaire testé sur des images comprimées possède également une fonctionnalité de mesure des volumes des nodules pulmonaires. Il est donc intéressant de tester la robustesse de cette fonctionnalité à la compression. C'est l'objet de l'étude préliminaire qui est présentée dans le paragraphe suivant.

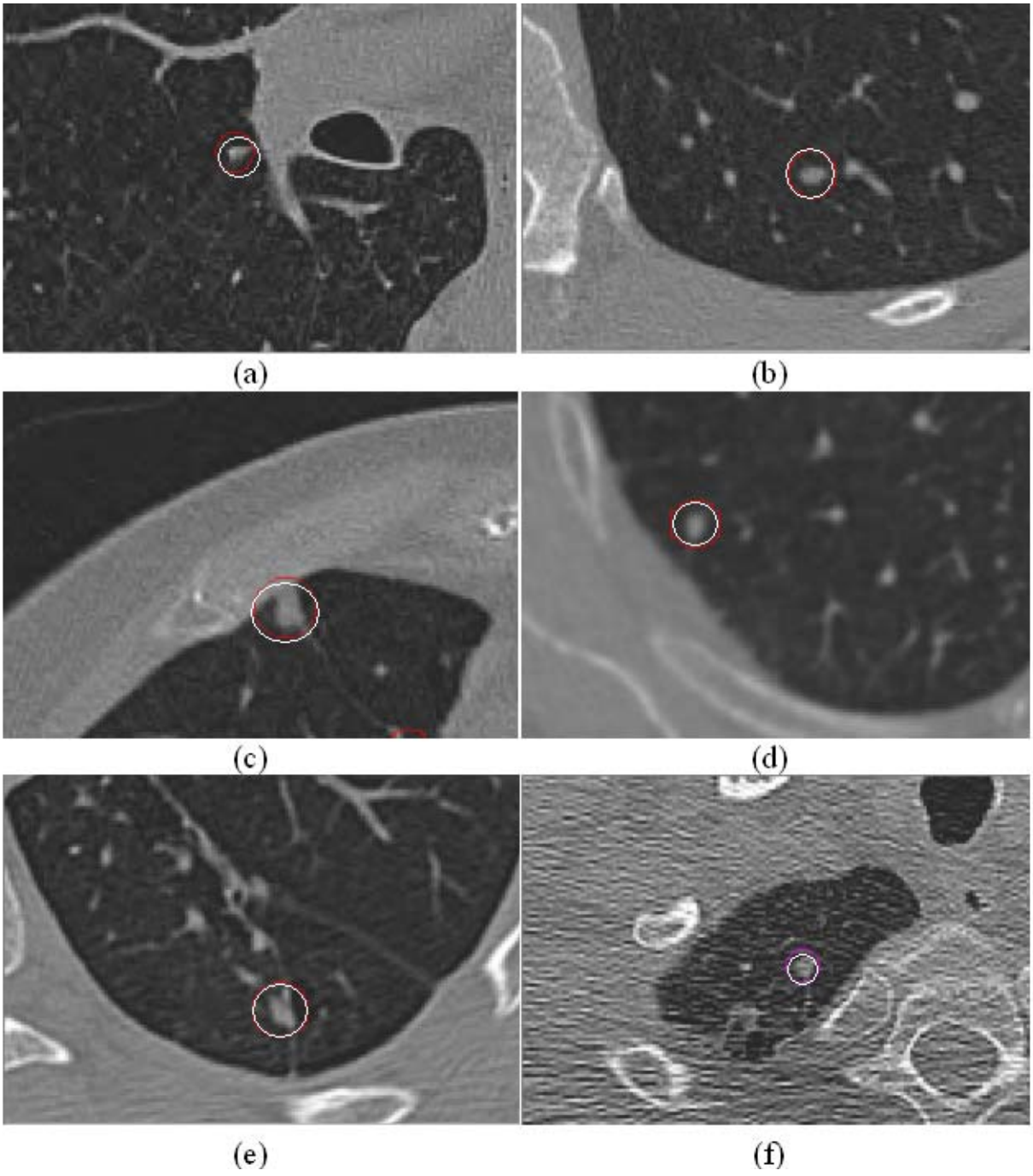

FIG. 3.3: Exemple de nodules manqués par le CAD sur des images non compressées et retrouvés à un taux de compression de 24 : 1 (cercles blancs). Les images sont affichées à 24 : 1. (a) Un nodule de 4,6 mm de taille touchant une fissure et à proximité du mediastinum (mA = 57, kVp = 120, noyau GE Lung , pas de media) (b) Nodule périphérique isolé de 4,3 mm (mA = 20, kVp = 140, noyau Siemens B60f, pas de media), (c) Nodule pleural de 6,2 mm (mA = 20, kVp = 140, noyau Siemens B60f , pas de media), (d) Nodule périphérique isolé de 4,7 mm (mA = 60, kVp = 120, noyau Siemens B30f, pas de media), (e) Nodule périphérique de 4,4 mm (mA = 175, kVp = 120, noyau GE Lung, IV pas de media), (f) Nodule apicale de 5,1 mm (mA = 59, kVp = 120, noyau GE Lung l, pas de media).

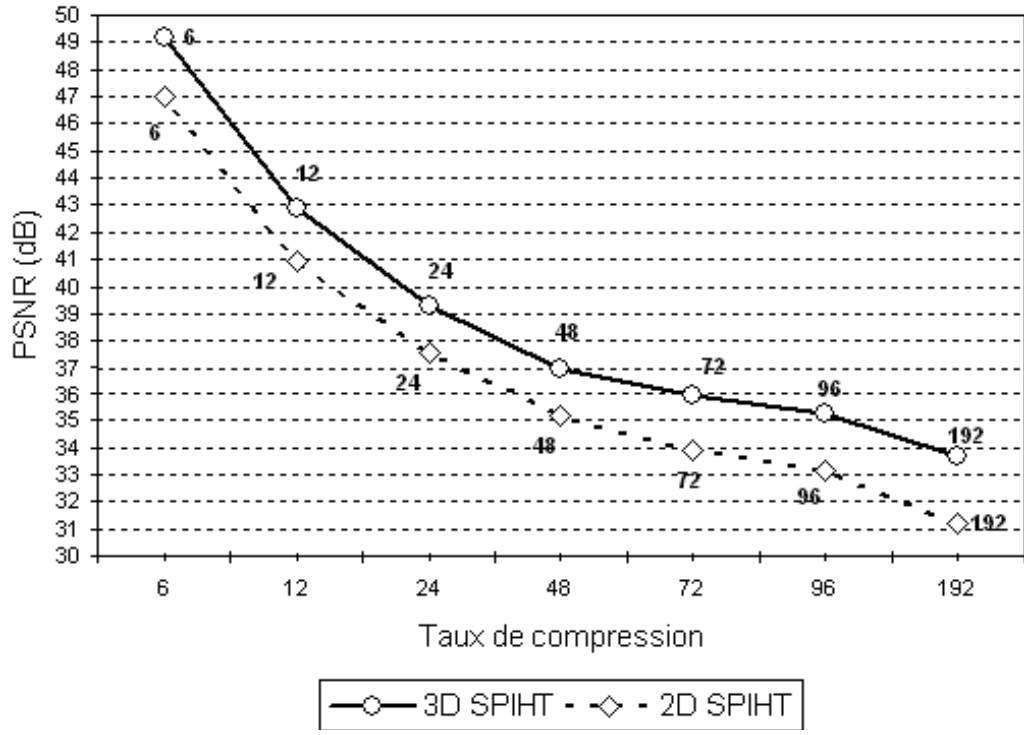

Fig. 3.4: Comparaison des performances de SPIHT 2D et SPIHT 3D en termes de moyenne des PSNR (en dB) moyen de 30 patients représentatifs à différents taux de compression.

3.5 Evaluation des effets de la compression sur la volumétrie

Le but ici est d'évaluer les effets de la compression avec pertes d'images volumiques sur les mesures de volume des nodules détectés par le CAD [91]. Les expériences ont été effectuées sur deux examens MCDT haute-résolution de poitrine : un normal et un pathologique (emphysème). Ils ont été acquis à 16 x 1,25 mm avec un intervalle de reconstruction de 0,6 mm. Dans ces deux examens, 89 nodules générés par ordinateur avec des volumes connus variant de 31,7 à 896,7 mm^3, reflétant des caractéristiques typiques de nodules solides d'examen MDCT de poitrine ont été insérés à des localisations au hasard, juxtavasculaires ou sous-pleurales. Les deux images 3D ont été comprimées en utilisant la méthode de compression SPIHT 3D aux même taux de compression que pour l'étude précédente (24 : 1, 48 : 1 et 96 : 1). Un modèle d'analyse de la variance (ANOVA) a été mis en place pour évaluer l'impact du taux de compression et des caractéristiques des nodules sur les estimations de volumes fournies par le CAD.

Le CAD produit une erreur mesurée moyenne sur le volume des nodules de + 3,4 % ± 7,6 % pour l'ensemble des données non compressées. Sur une image comprimée, par exemple à un taux de compression de 96 : 1, cette même erreur moyenne sur le volume est de + 0,5 % ± 9,1 % . En fait, aucune différence significative dans l'erreur de mesure des volumes pour tous les niveaux de compression n'a été identifiée. Cependant, bien que la compression ne puisse pas prévoir la taille des erreurs volumétriques, le taux de compression de 96 : 1 produit statistiquement des volumes sensiblement plus petits en moyenne (perte en moyenne de 4,1 mm^3 pour les volumes) que les autres taux de compression testés.

Par ailleurs, de manière identique à l'étude précédente sur la sensibilité du CAD, la localisation du nodule s'avère un facteur prédictif significatif d'erreur de mesure de volume avec de plus grandes erreurs en moyenne pour les nodules juxtavasculaires (+ 7,1 %) et pleuraux (+ 1.6 %). La taille des nodules, tout comme la présence d'emphysème n'ont pas été identifiées en tant que facteurs

prédictifs significatifs d'erreur de mesure de volume.

Ainsi, les évaluations de volume par le CAD des nodules sur des données comprimées avec pertes par SPIHT 3D n'ont montré aucune différence significative jusqu'à 96 : 1 en comparaison avec les volumes présents dans la base de données de référence.

3.6 Conclusion

Nous avons proposé deux résultats principaux sur des scanners de type MDCT de poumons. Le premier a montré que l**a détection des nodules par le CAD n'était pas détériorée par la compression basée sur SPIHT 3D jusqu'à un niveau de compression de 48 : 1 et restait robuste jusqu'à 96 : 1**. La seconde étude sur la volumétrie, bien que préliminaire, a déjà montré que les estimations **des volumes des nodules** sur des données **compressées avec pertes (par SPIHT 3D)** ne présentaient **aucune variation significative**, si l'on compare aux volumes connus dans la base de référence. Ces deux études font partie des premiers résultats pour l'évaluation des effets de la compression avec pertes en utilisant un codeur 3D sur les performances d'un CAD pulmonaire pour des images de type MDCT. Elles constituent une première validation pour l'utilisation éventuelle de méthodes de compression 3D pour les images MDCT des poumons.

Actuellement, la compression avec pertes est utilisée avec abondance (notamment sur Internet). En imagerie radiologie, les algorithmes 3D récents de compression avec pertes sont encore très loin d'être utilisés dans la pratique. Pour envisager un jour une mise en production de la compression avec pertes dans un PACS ou par les grands constructeurs d'appareils radiologiques, il est fondamental de **populariser l'idée de la compression avec pertes maitrisées dans la communauté médicale**. Multiplier les études validant les méthodes de compression avec pertes sur différents organes, modalités et traitements va clairement dans ce sens. Dans cette optique, nous réalisons actuellement une étude sur les effets de la compression avec pertes sur le recalage rigide et non rigide d'images oncologiques. De plus, nous débutons des tests sur l'impact des pertes liés à la compression sur des reconstructions volumiques classiques dans des scanners de reins. Une autre étude sur l'évaluation de segmentation cardiaque 3D d'images compressées est en projet.

Chapitre 4

Quelques outils analytiques pour la compression d'images

4.1 Introduction

L'approche intra-bande considère chaque sous-bande issue de la transformée en ondelettes de façon indépendante. Pour construire un schéma de compression efficace, il est alors nécessaire de déterminer un débit pour chacune d'elles à travers une procédure d'allocation de débits. Le second chapitre a présenté une nouvelle méthode de compression intra-bandes pour les images médicales volumiques basée sur une quantification algébrique avec zone morte. Ainsi, cette méthode nécessite une allocation de débits qui peut s'avérer très coûteuse en calculs, et ceci d'autant plus qu'elle requiert le réglage d'un paramètre supplémentaire : la zone morte vectorielle. L'objectif ici est de diminuer sensiblement le nombre de calculs effectués à partir des données. Nous proposons dans ce chapitre deux solutions pour l'allocation de débits de la QVAZM pour les images naturelles. La seconde solution est étendue aux images médicales 3D.

L'allocation de débits correspond à une minimisation non linéaire sous contraintes d'égalité. Elle a fait l'objet de nombreuses recherches, résumées dans [78]. Ces méthodes peuvent être découpées en trois groupes :

La première famille comprend les méthodes basées sur le calcul des courbes débit-distorsion. Deux approches peuvent être employées dans la résolution du problème de minimisation : l'approche lagrangienne ou la programmation dynamique [78]. Si celles-ci assurent la plus grande précision en terme de détection du minimum du fait de l'utilisation directe des données, elles présentent l'inconvénient d'un coût de complexité assez élevé. En effet, elles travaillent avec les courbes débit-distorsion qui nécessitent énormément de ressources pour être calculées. C'est ce type de méthode (lagrangienne) qui a été utilisé pour l'allocation du chapitre 2.

La deuxième classe de méthodes regroupe les approches du type approximation ou ajustement des courbes débit-distorsion, dont le principe consiste à effectuer un certain nombre de mesures sur les données (i.e. calculer quelques points de la courbe débit-distorsion) et à en déduire une approximation grâce à un modèle. Cette approche est un compromis entre nombre d'appels aux données et validité de l'estimation. Elle est particulièrement utilisée en compression vidéo pour le contrôle des débits [78], [17]. La méthode[1] proposée dans la première partie de ce chapitre pour l'allocation de la QVAZM 2D, peut être classée dans cette famille : elle consiste à remplacer l'appel aux données par une estimation des courbes débit-distorsion grâce à un modèle exponentiel [47], l'allocation en elle-même consistant à appliquer de simples formules analytiques.

Enfin, les approches statistiques ont pour intérêt de délivrer un modèle analytique dont la résolution est beaucoup moins coûteuse du fait de la quasi absence d'appel aux données dans le processus

[1] Ce travail a été effectué conjointement avec Mr Ludovic Guillemot, post-doctorant au CRAN.

d'allocation. De nombreux travaux ont porté sur des méthodes d'allocations basées sur des modèles statistiques des coefficients d'ondelettes [93], reposant sur des hypothèses de type gaussienne généralisée [80] ou faisant appel à des modélisations prenant en compte les dépendances spatiales [63]. Dans ce chapitre, nous présentons une nouvelle modélisation[2] de la norme des vecteurs en utilisant un mélange de lois gamma conduisant à un mélange de lois de type gaussienne généralisée des vecteurs de la source eux-mêmes [48]. Nous dérivons ainsi des modèles débit-distorsion pour la QVA et la QVAZM, simplement déduits des modèles classiques dans le cas i.i.d. Une fois la norme des vecteurs d'ondelettes modélisée par un mélange de gamma, nous ne faisons plus du tout appel aux données. La complexité de notre schéma de compression intra-bandes en est ainsi considérablement réduite.

Ce chapitre est composé de trois parties, la première présente la procédure analytique d'allocation de débits basée sur une approximation des courbes débit-distorsion de la QVAZM avec modèle exponentiel. La seconde partie propose des modèles statistiques débit-distorsion valables pour la QVA et la QVAZM. Ils sont dédiés à des mélanges de densité par blocs. La dernière partie compare les performances de ces allocations sur un panel classique d'images naturelles. Elle présente également les résultats de l'allocation avec le modèle statistique sur un scanner ORL de la base E3.

4.2 Allocation de débits pour la QVAZM par approximation des courbes débit-distorsion à l'aide d'un modèle exponentiel

Dans un premier temps, nous allons proposer un modèle pour les courbes débit-distorsion du schéma QVAZM, motivé d'une part par les théories de Shannon et haute résolution, et d'autre part, par l'expérimentation. Nous présenterons ensuite la démonstration de la résolution analytique de l'allocation des débits.

4.2.1 Détermination du modèle des courbes débit-distorsion

La théorie de Shannon et la théorie haute résolution délivrent des informations précieuses pour la détermination d'un modèle [45]. Nous allons rappeller brièvement ces deux théories.

1. **Théorie de Shannon :** A l'image de l'entropie d'une source, elle permet de donner une borne inférieure de la distorsion qu'il est possible d'atteindre à un débit donné. C'est une donnée asymptotique : cette borne est obtenue en faisant tendre la dimension de quantificateurs vectoriels vers l'infini. Une des restrictions majeures réside dans le fait que cette borne dépend de la statistique de la source : sa formule explicite n'est connue que dans le cas de distributions i.i.d. telles que la loi gaussienne ou laplacienne et les processus gaussiens [107].
2. **Théorie haute résolution** : Cette théorie a déjà été présentée dans le cadre de la QVA dans le deuxième chapitre. D'une manière générale, elle permet d'obtenir une approximation des courbes débit-distorsion pour un schéma de compression donné. Malheureusement, cette approximation n'est valide qu'à haut débit (lorsqu'on peut faire l'hypothèse d'un bruit de quantification uniforme).

Nous avons donc d'un côté une théorie qui permet d'obtenir une courbe "limite" débit-distorsion qui est toutefois assez difficile à déterminer de manière analytique, et d'un autre côté, la théorie haute résolution permettant d'obtenir une approximation mais uniquement dans la cas des forts débits. En combinant ces deux informations, on aboutit à une formule générale [51], [107] de la fonction distorsion en fonction du débit D pour une source de variance σ^2 :

$$D(R) = g(R)\,\sigma^2 2^{-2R} \tag{4.1}$$

[2] Ce travail a été effectué en collaboration avec Mr Ludovic Guillemot, post-doctorant au CRAN et Mr Saïd Moussaoui, Maître de Conférences à l'Ecole Centrale de Nantes.

avec $g \to c$ lorsque R tend vers l'infini et $g(0) = 1$.

c est une constante dépendant des statistiques de la source.

La fonction g étant inconnue, il faut déterminer un modèle permettant d'approximer g de manière satisfaisante.

Une solution consiste à interpréter g comme une "modulation" de la décroissance exponentielle en 2^{-2R} du modèle (4.1).

Dès lors, une solution consiste à considérer le modèle suivant :

$$D(R) = Ce^{-aR} \tag{4.2}$$

avec C et a les paramètres du modèle.

Le point clef de notre méthode par rapport aux méthodes employant une fonction exponentielle [96] réside dans le fait que l'ajustement s'effectue par l'intermédiaire du facteur de décroissance exponentielle a : on obtient ainsi un compromis entre décroissance exponentielle en 2^{-2R} lorsque R est grand et une décroissance généralement plus forte près de zéro. Soulignons enfin que certains travaux ont utilisé un modèle exponentiel pour modéliser les relations entre débit, distorsion et pas de quantification du flux MPEG [40].

L'autre argument en faveur de ce modèle concerne la résolution du problème d'allocation des débits que (4.2) permet de résoudre de manière analytique. Ce dernier point sera détaillé dans la deuxième partie de ce paragraphe.

4.2.1.1 Approximation

Pour une source donnée, il faut déterminer les paramètres C et a du modèle minimisant l'erreur entre la courbe expérimentale et la courbe estimée. Cette phase est fondamentale : de l'efficacité de l'ajustement dépend la validité de l'allocation des débits.

L'ajustement à partir d'un modèle exponentiel est équivalent à une régression linéaire. En effet on a :

$$\ln[D(R)] = \ln(C) - aR \tag{4.3}$$

La régression linéaire [84] est un cas particulier du problème des moindres carrés :

$$\min_P \|(XP - Y)\|^2 \tag{4.4}$$

où X est la matrice dépendant du modèle, $P = (p_1, ..., p_n)$ le vecteur des paramètres à déterminer et $Y = (y_1, ..., y_L)$ les mesures.

Dans le cas de la régression linéaire on obtient les formules explicites suivantes :

$$\begin{aligned} p_1 &= \frac{L\sum_{i=1}^{L} x_i y_i - \sum_{i=1}^{L} x_i \sum_{i=1}^{L} y_i}{L\sum_{i=1}^{L} x_i^2 - \left(\sum_{i=1}^{L} x_i\right)^2} \\ p_2 &= \frac{1}{L}\left(\sum_{i=1}^{L} y_i - p_1 \sum_{i=1}^{L} x_i\right) \end{aligned} \tag{4.5}$$

Par conséquent, en appliquant (4.5) à (4.3), on obtient :

$$\begin{aligned} C &= \exp\left(\frac{L\sum_{i=1}^{L} R_i d_i - \sum_{i=1}^{L} R_i \sum_{i=1}^{L} d_i}{L\sum_{i=1}^{L} R_i^2 - \left(\sum_{i=1}^{L} R_i\right)^2}\right) \\ a &= -\frac{1}{L}\left(\sum_{i=1}^{L} d_i - \ln(C)\sum_{i=1}^{L} R_i\right) \end{aligned} \tag{4.6}$$

avec $d_i = D(R_i)$ la mesure de la distorsion pour le débit R_i.

Nous venons de présenter la forme la plus classique de la méthode des moindres carrés. Cependant certaines particularités des courbes distorsion-débit sont à prendre en compte :

1. **Fortes pentes :** Une petite erreur dans une zone de forte pente peut entraîner une grande différence au niveau des débits alloués. En revanche une petite erreur dans les forts débits entraînera de faibles différences, les pentes y étant beaucoup plus faibles.
2. **A priori sur les débits :** Un schéma de compression par transformée en ondelettes privilégiera toujours les basses résolutions : de forts débits seront donc alloués à ces sous-bandes. Inversement les débits les plus faibles correspondent aux grandes sous-bandes. Il s'agit là d'un *a priori* sur les valeurs des débits alloués en fonction du niveau de résolution qu'il est possible d'exploiter au niveau du choix des mesures : leur répartition est déterminée en fonction du niveau de résolution de la sous-bande (une solution alternative consiste à employer les moindres carrés pondérés, la pondération permettant de minimiser l'erreur d'approximation dans la zone privilégiée).

4.2.1.2 Résultats expérimentaux

La figure 4.1 permet de comparer la courbe débit-distorsion obtenue par l'approximation exponentielle à celle calculée à partir des données réelles sur la sous-bande diagonale de niveau 3 (D3)[3] de l'image de Lena (4.2). L'approximation a été obtenue en calculant la distorsion pour seulement trois débits (en bits par échantillon) : 0,2 ; 0,4 ; 0,6. Comme on peut le voir, l'approximation de la courbe est particulièrement efficace sur la plage de débit $[0, 1]$. Au delà, le modèle a tendance à diverger mais cela concerne les hauts débits qui ont peu de chance de constituer le débit cible dans notre schéma de compression.

Ceci est vérifié par la figure 4.2 qui compare le logarithme (népérien) de la distorsion en fonction du débit avec celui du modèle exponentiel. Comme on peut le voir sur cette figure, le modèle diverge pour les grands débits. Cela illustre l'importance du choix des points sur lesquels vont s'effectuer l'approximation : on privilégiera ainsi les faibles débits pour les grandes sous-bandes et les grands débits en ce qui concerne les petites sous-bandes. La figure 4.3 permet de montrer la validité de notre modèle dans la sous-bande V1 de l'image Lena sur la plage de débits qui concerne cette sous-bande.

Pour conclure, il est important de souligner qu'à ce stade, cette étude a constitué une première solution pour améliorer l'allocation de la QVAZM. Celle-ci souffre de l'absence de formalisation, à savoir de quelle manière établir le compromis entre approximation locale et approximation globale. La recherche d'un modèle statistique plus fidèle est développée dans le paragraphe 4.3. Néanmoins, nous verrons à la fin du chapitre que le modèle exponentiel conduit à une allocation dont les résultats sont assez proches d'une méthode de type gradient que nous avons introduite dans le chapitre 2.

4.2.2 Allocation des débits

La résolution du problème d'optimisation va s'effectuer grâce à l'approche lagrangienne, dite des pentes égales ("equal slope" en anglais) dont une présentation claire est proposée dans [107] dans le cas d'un modèle analytique proche de la fonction définie dans la formule (4.1).

Après avoir rappelé la formalisation du problème général puis dans le cas de fonctions exponentielles, nous expliciterons la solution du problème. Les résultats expérimentaux seront présentés à la fin du chapitre. La figure 4.4 résume le processus d'allocation de débits utilisant le modèle exponentiel.

[3] A partir d'ici, nous proposons une écriture simplifiée pour les sous-bandes produites par une TO2D dyadique. Les lettres V, H et D signifient respectivement une sous-bande verticale, horizontale et diagonale. Le chiffre suivant la lettre correspond au niveau de décomposition.

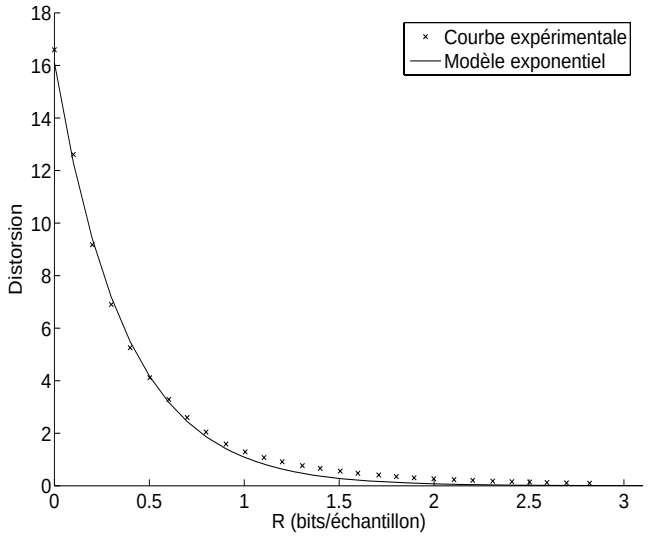

Fig. 4.1: Comparaison des courbes de distorsion en fonction du débit obtenues expérimentalement et à partir du modèle exponentiel (construit à partir de 3 points de débit 0,2 ; 0,4 et 0,6 bits par échantillon) - sous-bande D3 de l'image Lena.

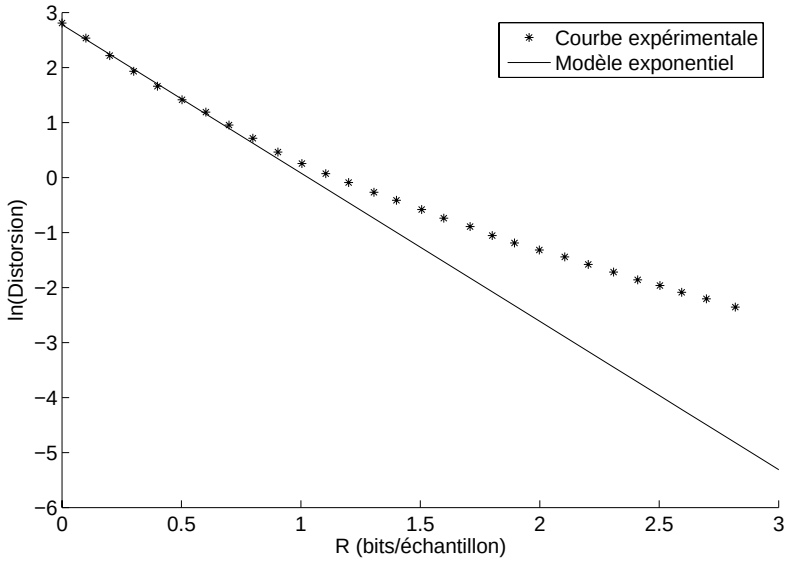

Fig. 4.2: Comparaison des logarithmes (népériens) de la courbe expérimentale et du modèle (construit à partir de 3 points de débit 0,2 ; 0,4 et 0,6 bits par échantillon) - sous-bande D3 de l'image Lena.

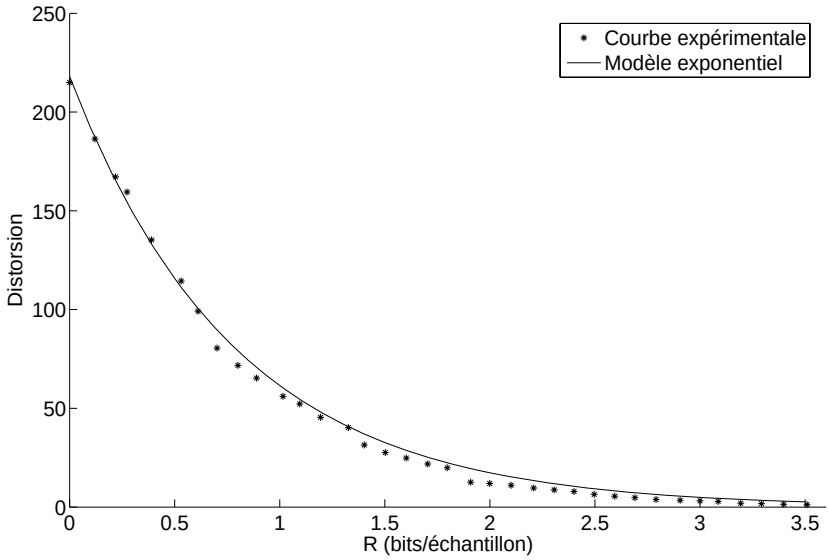

FIG. 4.3: Comparaison des courbes de distorsion en fonction du débit obtenues expérimentalement et à partir du modèle exponentiel (construit à partir de 3 points de débit 0,2 ; 0,4 et 0,6 bits par échantillon) - sous-bande V1 de l'image Lena.

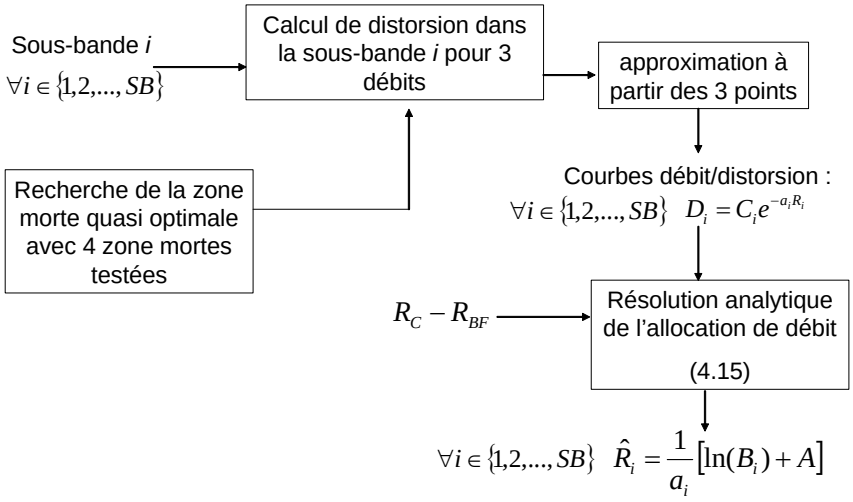

FIG. 4.4: Processus d'allocation de débits utilisant une approximation par un modèle exponentiel des SB courbes $D_i(R_i)$.

4.2.2.1 Formalisation du problème

Nous rappelons que le problème d'allocation des débits pour une analyse multirésolution de SB sous-bandes (en excluant encore une fois de l'allocation la sous-bande basse-fréquence) est le problème de minimisation suivant :

$$\begin{cases} \widehat{R}_T = \arg\min_{(R_1,...,R_{SB})} \left(\sum_{i=1}^{SB} \beta_i D_i(R_i) \right) \\ \sum_{i=1}^{SB} \alpha_i R_i = R_C \end{cases} \tag{4.7}$$

avec D_i les fonctions débit-distorsion de chaque sous-bande, R_C le débit cible à atteindre, α_i le rapport entre le nombre d'échantillons dans la sous-bande i et le nombre total de pixels de l'image. Les facteurs β_i dépendent de la non-orthogonalité des filtres et d'éventuelles pondérations psychovisuelles. $\widehat{R}_T = \left(\widehat{R}_1, ..., \widehat{R}_{SB}\right)$ est le vecteur solution.

Classiquement le problème (4.7) se résoud grâce à l'approche lagrangienne [11] [51] [78] [80] [107]qui permet de s'affranchir de la contrainte d'égalité en incluant celle-ci dans la fonctionnelle afin d'obtenir le problème de minimisation suivant :

Soit $\widehat{R}_T = \left(\widehat{R}_1, ..., \widehat{R}_{SB}\right)$ la solution du problème (4.7) alors on a :

$$\exists \lambda > 0 / \ R = \arg \min_{(R_1,...,R_{SB})} \left(\sum_{i=1}^{SB} \beta_i D_i \left(R_i\right) + \lambda \left(\sum_{i=1}^{SB} \alpha_i R_i - R_C \right) \right) \tag{4.8}$$

λ est appelé le multiplicateur de Lagrange.

Par conséquent, en calculant le gradient on obtient les SB égalités suivantes :

$$\frac{\partial}{\partial R_i} \left(\beta_i D_i \left(R_i\right) + \lambda \left(\alpha_i R_i\right) \right) = 0, \ \forall i = 1, ..., SB \tag{4.9}$$

On en déduit les égalités suivantes :

$$\lambda = -\frac{\beta_i}{\alpha_i} D_i^{'} \left(R_i\right) \quad \forall i = 1, ..., SB \tag{4.10}$$

où $D_i^{'}$ correspond à la dérivée de D_i par rapport à R_i.

Dans la suite, nous allons proposer une résolution analytique au problème (4.7) à partir du modèle établi dans le paragraphe précédent.

4.2.2.2 Résolution analytique

Nous supposons que les fonctions $D_i\left(.\right)$ sont définies analytiquement par :

$$D_i \left(R_i\right) = C_i e^{-a_i R_i} \tag{4.11}$$

avec C_i et a_i appartenant à $\mathbb{R}^+$.

On obtient le problème suivant :

$$\begin{cases} \widehat{R}_T = \arg\min_{(R_1,...,R_{SB})} \left(\sum_{i=1}^{SB} \beta_i C_i e^{-a_i R_i} \right) \\ \sum_{i=1}^{SB} \alpha_i R_i = R_C \end{cases} \tag{4.12}$$

Dans le cas de fonctions de type (4.11), l'égalité (4.10) permet d'obtenir une formule explicite des composantes du vecteur solution $\widehat{R}_T$: on a $\lambda = C_i \frac{\beta_i}{\alpha_i} a_i e^{-a_i R_i}$, d'où :

$$R_i = \frac{1}{a_i} \ln \left(\frac{C_i \beta_i a_i}{\lambda \alpha_i} \right) \tag{4.13}$$

On a ainsi exprimé les débits R_i en fonction de λ.
A partir de la contrainte d'égalité du problème (4.12), on obtient :

$$R_C = \sum_{i=1}^{SB} \alpha_i R_i = \sum_{i=1}^{SB} \mu_i \ln \left(\frac{C_i \beta_i a_i}{\lambda \alpha_i} \right)$$

avec $\mu_i = \frac{\alpha_i}{a_i}$
D'où $R_C = \sum\limits_{i=1}^{SB} \ln \left(\left(\frac{C_i \beta_i a_i}{\lambda \alpha_i} \right)^{\mu_i} \right) = \ln \left(\prod\limits_{i=1}^{SB} \left(\frac{C_i \beta_i a_i}{\lambda \alpha_i} \right)^{\mu_i} \right)$
Ce qui conduit à une formule explicite du multiplicateur de Lagrange λ :

$$\lambda = \left(\frac{\prod\limits_{i=1}^{SB} (B_i)^{\mu_i}}{e^{R_C}} \right)^{\frac{1}{\sum\limits_{l=1}^{SB} \mu_l}} \tag{4.14}$$

avec $B_i = \frac{C_i \beta_i a_i}{\alpha_i}$.
En injectant (4.14) dans (4.13) on obtient la solution analytique suivante :

$$\begin{aligned} \widehat{R}_i &= \frac{1}{a_i} \left[\ln (B_i) + A \right] \\ A &= \frac{1}{\sum\limits_{l=1}^{SB} \mu_l} \left[R_C - \sum_{i=1}^{SB} \mu_i \ln (B_i) \right] \end{aligned} \tag{4.15}$$

L'allocation des débits devient donc une simple formule analytique dont le coût de calcul est dérisoire. La complexité de l'allocation se limite aux calculs des trois points utiles pour l'approximation des courbes débit-distorsion par le modèle exponentiel.

Remarque : Si l'on souhaite calculer l'erreur quadratique moyenne prédite par le modèle, celle-ci s'obtient en injectant (4.15) dans les formules données par (4.11) : $D_i(R_i) = C_i e^{-a_i R_i} = \frac{\mu_i}{\beta_i} e^{-A}$. D'où :

$$D(R) = \sum_{i=1}^{SB} \beta_i D_i(R_i) = e^{-A} \sum_{i=1}^{SB} \mu_i \tag{4.16}$$

4.2.2.3 Complexité de l'allocation de débits

Nous allons exprimer la complexité C_{exp} de l'allocation de débits correspondant à la figure 4.4. Pour déterminer les paramètres du modèle dans chaque sous-bande i, notre méthode a donc seulement besoin de calculer trois points débit-distorsion. L'approximation a été obtenue en calculant la distorsion de trois débits cibles. Ainsi, nous appelons trois fois l'algorithme rapide de recherche de la zone morte (qui minimise la fonction à une seule variable $D^*(r_{ZM})$) présentée dans le chapitre 2. Pour minimiser $D^*(r_{ZM})$, cette méthode proche de l'optimalité réduit la recherche de la zone morte à une séquence de test de valeurs de rayon du type $r_l = l - \varepsilon$ avec $l \in \mathbb{N}$ et $\varepsilon \ll 1$. Pratiquement,

en choisissant $\varepsilon = 0,1$, limiter la recherche de la zone morte à un ensemble de quatre zones mortes ($r_0 = 0\,;\ r_2 = 1,9\,;\ r_3 = 2,9$ et $r_4 = 3,9$) est suffisant pour obtenir une minimisation proche de l'optimalité. De plus, pour chaque zone morte testée, la recherche du facteur de projection γ^* permettant d'atteindre le débit cible est réalisée. On rappelle que cette recherche utilise une méthode dichotomique dont le nombre d'itérations, c'est à dire la rapidité, est lié à la grandeur de l'intervalle de recherche de $\gamma^*(.)$. Le chapitre 2 a présenté une première relation pour réduire de manière significative le domaine de recherche de γ^*. En effet, si l'on considère la suite $(r_l)_{l\in N}$ définie plus haut, nous avons montré l'encadrement suivant :

$$\frac{r_l}{r_{l+1}}\gamma^*(r_l) < \gamma^*(r_{l+1}) < \gamma^*(r_l) \tag{4.17}$$

De plus, Il est également possible d'ajouter d'autres encadrements liés au fait que l'on calcule des points R-D à des débits croissants. ($R_1 < R_2 < R_3$). Pour une même zone morte r_l donnée, on a l'inégalité suivante :

$$\gamma^*_{R_3}(r_l) < \gamma^*_{R_2}(r_l) < \gamma^*_{R_1}(r_l) \tag{4.18}$$

En ajoutant d'autre encadrements s'appuyant sur l'importance des sous-bandes (à travers l'énergie d'une sous-bande) quatre ou cinq itérations suffisent en moyenne à trouver γ^*.

L'ensemble de ces remarques conduisent à la proposition suivante :

Proposition 2 *La complexité calculatoire $C_{\exp}$ de l'allocation de débits de la QVAZM à partir du modèle exponentiel peut être exprimée approximativement par*

$$C_{\exp} \simeq \sum_{i=1}^{SB} \alpha_i \left(36 + 36 \times \#I_i\right) (op/pixel) \tag{4.19}$$

où op désigne une opération élémentaire ($+$, $-$, $\times$, $\div$) et $\#I_i$ est le nombre d'itérations moyen pour trouver le minimum de $D^*(r_{ZM})$

Preuve

Nous considérons uniquement les opérations élémentaires *op* permettant le calcul des paramètres du modèle. Certaines opérations considérées comme négligeables sont exclues de l'évaluation :

- l'opération d'arrondi de la quantification,
- le calcul du débit théorique,
- la résolution analytique de la minimisation une fois les fonctions exponentielles estimées.

A partir de ces exclusions, nous pouvons approximer la complexité pour calculer les paramètres par l'équation suivante :

$$C_{\exp} \simeq \sum_{i=1}^{SB} \alpha_i \times \#_{pts} \times \#_{ZM} \times \left(\#_{dist} + \#_{quant} \times \#I_i\right)\ (op/pixel) \tag{4.20}$$

où $\#_{pts}$ est le nombre de points nécessaires pour approximer le modèle, $\#_{ZM}$ est le nombre de zones mortes testées pour la minimisation de $D^*(r_{ZM})$, $\#_{dist}$ représente le nombre d'opérations pour calculer la distorsion d'un pixel et $\#_{quant}$ correspond au nombre d'opérations pour la quantification d'un pixel. Dans le cas de notre schéma de compression, les valeurs citées ci-dessus sont les suivantes :

- $\#_{quant} = 3$ car nous avons trois opérations liées à la quantification pour un pixel : une division correspondant à la projection par γ, une multiplication par $\frac{1}{\gamma}$ qui est l'opération de quantification inverse et une addition dans le calcul la norme L_1 du vecteur.

- $\#_{dist} = 3$ car il y a trois opérations liées au calcul de la distorsion : une soustraction entre le pixel orignal et celui reconstruit, une multiplication pour élever au carré et une addition dans le calcul de la distorsion totale.
- $\#_{pts} = 3$ car trois points suffisent pour une bonne précision du modèle exponentiel
- $\#_{ZM} = 4$. En effet, on limite la recherche de la zone morte optimale minimisant $D^*(r_{ZM})$ à quatre zones mortes, comme on l'a dit au début du paragraphe.

Enfin, pour une allocation à bas débit (par exemple 0,125 bit/pixel ou 0,25 bit/pixel), certaines des plus grandes sous-bandes (typiquement V1, H1 ou D1) auront souvent un débit alloué égal à zéro. En faisant un tel *a priori* en fonction de l'énergie de ces grandes sous-bandes, nous pouvons encore réduire la complexité calculatoire. Par exemple, pour l'image boat à un débit inférieur ou égal à 0,125 bit/pixel, C_{exp} tombe typiquement à 45 opérations/pixel.

L'allocation proposée dans ce paragraphe est basée sur un ajustement des courbes débit-distorsion en les approximant par un simple modèle exponentiel. Les paramètres de ce modèle sont calculés à partir d'un nombre réduit d'appels aux données. Ces courbes exponentielles conduisent à une résolution analytique de l'allocation qui ne rentre pas en compte dans la complexité calculatoire totale. Néanmoins, le modèle exponentiel proposé ici manque un peu de flexibilité pour certaines sources dont les courbes débit-distorsion ne suivent pas une simple fonction exponentielle à deux paramètres (comme certaines sous-bandes d'une TO3D d'une image médicale). Nous proposons dans le paragraphe suivant une allocation de débits utilisant une approche statistique. Elle est basée sur des modèles débit-distorsion dédiés aux mélanges de densités par blocs.

4.3 Modèles débit-distorsion dédiés aux mélanges de densités par blocs

Nous proposons dans ce paragraphe des modèles débit-distorsion dédiés à la QVA/QVAZM qui se basent sur un mélange de densités par blocs [48]. Tout d'abord, nous montrons que la propriété d'agglutinement des signaux creux et structurés comme les coefficients conduit à une distribution de l'énergie des blocs de coefficients d'ondelettes proche d'un mélange de densités gamma. De ce modèle, la distribution conjointe des vecteurs est déduite. Elle correspond à un Mélange de densités Gaussienne Généralisée Multidimensionnelle (MMGG[4]). Le modèle MMGG permet de calculer un modèle statistique plus précis pour la QVA et donne un outil théorique pour la conception de quantificateurs mieux adaptés aux données comme par exemple un quantificateur algébrique avec zone morte. Nous déduisons finalement un modèle fidèle pour la QVAZM. Ce modèle est utilisé ensuite pour l'allocation de débits.

4.3.1 Introduction

De nombreux travaux dans le champ de la modélisation des courbes débit-distorsion pour la QVA à partir des caractéristiques statistiques des vecteurs ont été proposés (par exemple [18], [53]). Cependant, ces modèles font tous l'hypothèse que les signaux sont indépendants et identiquement distribués (i.i.d) alors que des méthodes de traitement du signal orientées blocs (comme la QVA) peuvent tirer profit des propriétés non i.i.d des données réelles (comme l'agglutinement des coefficients d'ondelettes par exemple). Dans ce domaine, Fisher *et al* ont proposé l'utilisation de dictionnaires elliptiques [38]. Cependant, cette approche nécessite une très grande complexité puisque elle requiert le calcul de l'orientation elliptique et l'application d'une méthode spéciale

[4]MMGG correspond aux initiales de l'acronyme anglais : Mixture of Multidimensional Generalized Gaussian densities.

d'indexage [74]. De plus, ces schémas exploitent des polarisations entre les échantillons qui correspondent seulement à une sorte de dépendances entre les échantillons.

Le but ici est de prendre en compte les effets d'agglutinement à l'intérieur des données. L'idée principale de l'approche proposée consiste à modéliser les dépendances à l'intérieur des vecteurs en utilisant un modèle statistique original sur la fonction de distribution de probabilité conjointe des vecteurs de la source. Ce modèle consiste en un modèle MMGG déduit d'un mélange de densités gamma pour la distribution de la norme. Cette approche présente deux apports principaux. Premièrement on déduit un nouveau modèle R-D (appelé modèle de mélange R-D) plus fidèle à la statistique de la source qui sera un outil efficace pour notre problème d'allocation de débits. Il surpasse les modèles R-D basés sur l'hypothèse i.i.d et peut être vu comme une généralisation des modèles non asymptotiques proposés par Raffy *et al* [89] pour la QVA. Deuxièmement, le modèle MMGG fournit une justification théorique à l'utilisation d'un dictionnaire adapté à la source comme celui employé dans la QVAZM.

4.3.2 Modélisation de la norme des vecteurs et de la source en utilisant un mélange de densités par bloc

Dans le paragraphe (2.3 du chapitre 2) sur l'orientation des vecteurs, nous avons vu que la norme des vecteurs pouvait être considérée comme une mesure de l'activité locale du signal. Dans cet exemple, on utilisait la norme L^1 pour les vecteurs qui représente la moyenne (en valeur absolue) des coordonnées du vecteur. Ainsi, dans le cas de signaux creux et agglutinés comme les coefficients d'ondelettes, la distribution de la norme dépend principalement des corrélations entre les échantillons. Par conséquent, comme l'efficacité d'un schéma de QVA sur une source non uniforme est principalement liée au codage entropique des vecteurs, les performances de codage sont le résultat des dépendances entre les échantillons. Nous appellerons cela le gain de parcimonie. Notons que celui-ci est différent du gain classique de corrélation qui est exploité par des dictionnaires elliptiques [38], puisque cela concerne la magnitude des coefficients et non les polarisations.

Pour rappeler cet aspect, nous faisons une comparaison entre les histogrammes de la norme de trois sortes de vecteurs : Les figures 4.5 (A) et (B) correspondent aux vecteurs de la sous-bande V3 de l'image de Lena avec deux orientations différentes : verticale (8×1) et horizontale (1×8) respectivement. La figure 4.5 (C) correspond à des vecteurs de forme 4×2 d'une source synthétique i.i.d, avec le même écart type ($\sigma = 7.7$) que la sous-bande considérée. Trois remarques peuvent être déduites de la figure 4.5 :

1. Les distributions empiriques (A) et (B) ont un mode proche de zéro. Ce mode est dû au voisinage des coefficients de faible valeur qui conduit à un grand nombre de vecteurs avec une petite norme.
2. La distribution (A) est plus piquée que la (B) puisque des vecteurs avec la même orientation que les détails capturent mieux les voisinages de coefficients significatifs et, inversement augmentent la concentration des vecteurs autour du vecteur nul.
3. La distribution i.i.d laplacienne sur la figure 4.5 (C) est très différente de celle de la sous-bande puisque son mode est éloigné de zéro (maximum en $\frac{\sigma(n-1)}{\sqrt{2}} = 38,1 >> 0$). Ainsi, le modèle i.i.d laplacien ne permet pas de représenter ce type de source.

Dans la plupart des travaux, la distribution de la norme est déduite de la distribution des échantillons de la source, que l'on suppose souvent i.i.d. Cette hypothèse conduit à des développements analytiques simples mais ne donne pas une description correcte des propriétés de la norme (et évidemment ne prend pas en compte les dépendances à l'intérieur des blocs de la source). En effet, si on fait l'hypothèse d'une distribution sur les échantillons suivant une loi gaussienne généralisée, la valeur du mode est seulement liée à l'écart-type. Ainsi, un grand écart-type conduira à un mode éloigné de zéro et entrainera un modèle non fiable pour la norme. En supposant qu'un

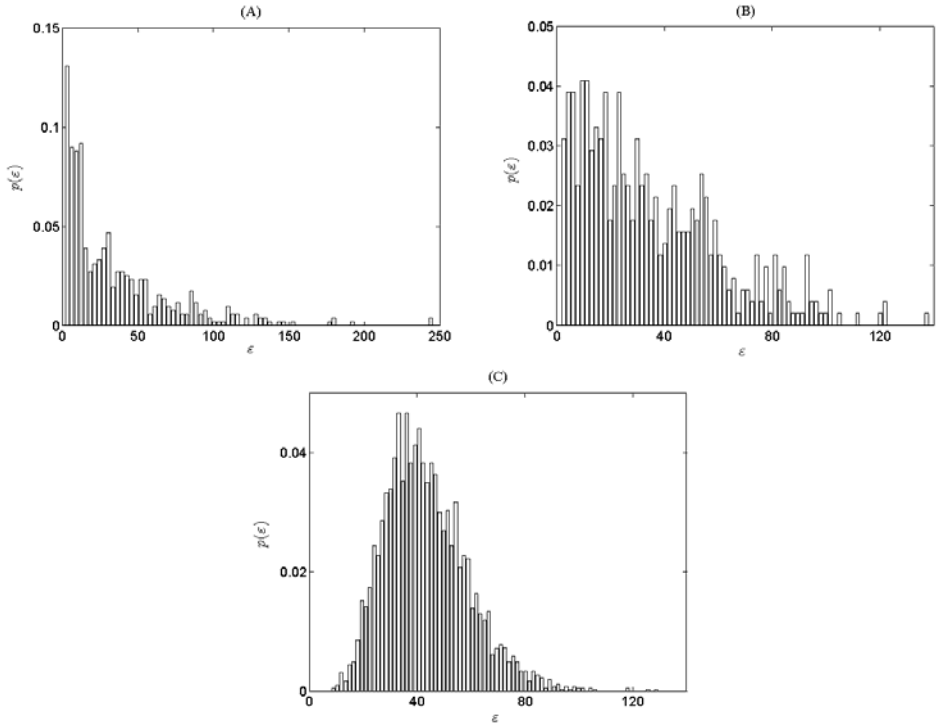

FIG. 4.5: Histogrammes de la norme L^1 pour : des vecteurs verticaux (A), des vecteurs horizontaux (B) de la sous-bande V3 de Lena et (C) des vecteurs distribués i.i.d laplacien (orientation : 4×2).

modèle statistique réaliste de la norme donne une description précise de celle-ci, nous proposons une approche inverse : nous déterminons un modèle flexible pour la norme qui permettra de déduire la distribution conjointe de la source des vecteurs.

4.3.2.1 Modélisation de la norme en utilisant un modèle de mélange de lois gamma

L'histogramme de la distribution de la norme de la figure 4.5, avec un mode proche de zéro et une longue queue nous amène à tester un modèle consistant en un mélange de densités. De plus, comme la norme des vecteurs est non négative, le mélange de lois gamma semble un bon candidat.

Nous rappelons qu'une variable aléatoire ε qui est distribuée suivant une loi gamma (dans la suite ε représente la norme des blocs) a la densité de probabilité suivante :

$$\mathcal{G}(\varepsilon; a, b) = \frac{b^a}{\Gamma(a)} \varepsilon^{a-1} \exp[-b\,\varepsilon]\ \mathbb{I}_{\varepsilon>0}, \tag{4.21}$$

où $\Gamma(a)$ est la fonction gamma, $\mathbb{I}_{\varepsilon>0}$ est la fonction indicatrice et les paramètres $(a > 0, b > 0)$ permettent d'ajuster la forme de la densité gamma (noté $\mathcal{G}(\varepsilon; a, b)$). Conne nous pouvons le remarquer sur la figure 4.6, pour $0 < a < 1$ la distribution est pointue et bien adaptée à des signaux parcimonieux ; pour $a > 1$, la distribution possède un mode en $(a-1)/b$ et est très allongée.

En supposant que la norme des vecteurs ε appartient à N_{mix} classes S_k de poids $c_k = P(\varepsilon \in S_k)$, nous obtenons le modèle de mélange suivant :

$$p(\varepsilon; \mathbf{a}, \mathbf{b}) = \sum_{k=1}^{N_{mix}} c_k\, \mathcal{G}(\varepsilon; a_k, b_k), \tag{4.22}$$

avec $\mathbf{a} = [a_1, ..., a_{N_{mix}}]$, $\mathbf{b} = [b_1, ..., b_{N_{mix}}]$. Dans le cadre de notre application, la norme des vecteurs est supposée appartenir à **deux classes notées S_1 et S_2** : l'état **des normes de faible valeur**

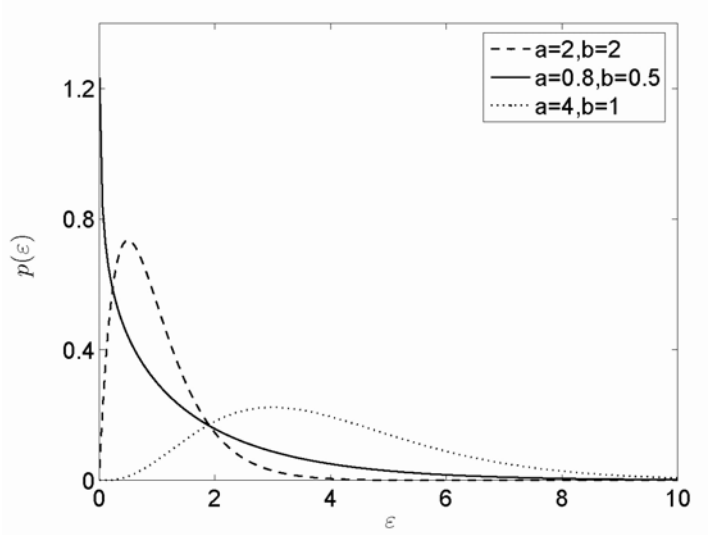

Fig. 4.6: Formes typiques de distributions gamma (équation 4.21).

et l'état correspondant **aux normes de grande valeu**r. Les paramètres (c_1, c_2) sont estimés en utilisant une simulation Monte-Carlo par chaîne de Markov (MCMC) [81]. Les paramètres des lois gamma **a** et **b** sont simulés en utilisant les algorithmes donnés dans [76]. Par exemple, nous testons ce modèle de mélange sur la sous-bande de l'image de Lena précédemment considérée (V3). La figure 4.7 montre que le mélange de deux lois gamma approxime bien la distribution de la norme L^1, c'est à dire un pic proche de zéro et une longue queue. La précision du modèle souligne le fait qu'une distribution conjointe conduisant pour la norme à une distribution de mélange de gamma semble être un bon représentant pour la distribution de la source elle-même.

4.3.2.2 Détermination de la distribution de la source

A partir de la distribution de la norme des vecteurs introduite en (4.22), la distribution conjointe f d'un vecteur aléatoire X représentant la source est donnée par :

$$f(X) = \sum_{k=1}^{N_{mix}} c_k f\left(X | \|X\|_\alpha^\alpha \in S_k\right) \tag{4.23}$$

où $f\left(X | \|X\|_\alpha^\alpha \in S_k\right)$ est la distribution du vecteur aléatoire X conditionnellement à l'état S_k.

Le choix de la distribution suivante $f\left(X | \|X\|_\alpha^\alpha \in S_k\right)$ est motivé par la proposition suivante :

Proposition 3 *Si nous considérons un ensemble de variables aléatoires $X = (x_1, ..., x_n)$, distribuées suivant une loi gaussienne généralisée, la norme L^α à la puissance α, $\varepsilon = \sum_{j=1}^{n} |x_j|^\alpha$, est distribuée suivant une loi gamma avec les paramètres $a = n/\alpha$ et $b = 1/\beta^\alpha$, où α et β sont respectivement la puissance et la forme de la densité de la gaussienne généralisée, qui s'écrit :*

$$\mathcal{GG}(x_i; \alpha, \beta) = \frac{\alpha\, \beta^{1/\alpha}}{2\Gamma(1/\alpha)} \exp(-\beta |x_i|^\alpha) \tag{4.24}$$

Preuve. Comme la distribution de chaque variable $x \in \{x_j; j = 1, ..., n\}$ est donnée par

$$p_X(x) = \frac{\alpha}{2\beta} \frac{1}{\Gamma(1/\alpha)} \exp\left(-\frac{|x|^\alpha}{\beta}\right)$$

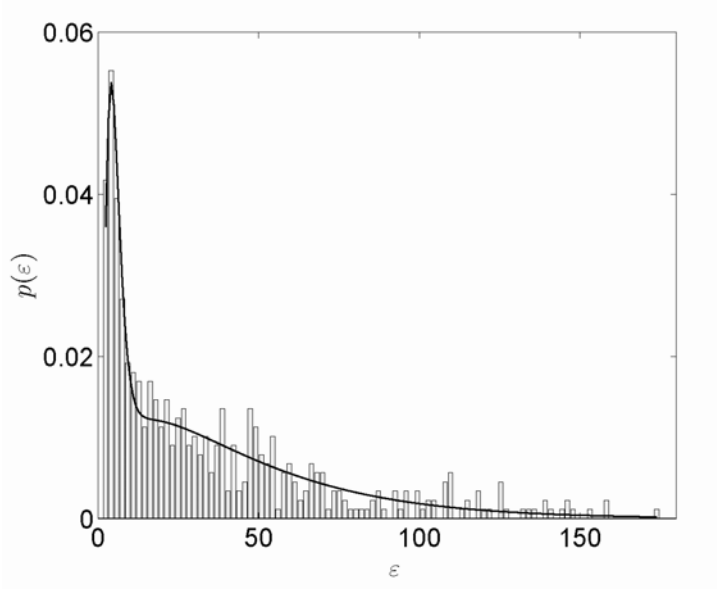

FIG. 4.7: Histogramme de la norme L^1 des vecteurs (2×4) de la sous-bande V3 de Lena et la densité correspondante d'un mélange de deux lois gamma estimées avec les paramètres $\mathbf{a} = [5,12 \quad 1,49]$, $\mathbf{b} = [0,92 \quad 0,03]$ et $\mathbf{c} = [0,24 \quad 0,76]$.

alors la distribution de sa valeur absolue est

$$p_{|X|}(|x|) = \frac{\alpha}{\beta}\frac{1}{\Gamma(1/\alpha)}\,\exp\left(-\frac{|x|^\alpha}{\beta^\alpha}\right)$$

La distribution de $v = |x|^\alpha$ est déduite selon

$$p_V(v) = p_X(x)\Big|_{|x|=v^{1/\alpha}}\,\left(\frac{dv}{d|x|}\right)^{-1}$$

où $\dfrac{dv}{d|x|} = \alpha\, v^{(\alpha-1)/\alpha}$.

Ainsi,

$$p_V(v) = \frac{1}{\beta}\frac{1}{\Gamma(1/\alpha)}v^{1/\alpha-1}\,\exp\left(-\frac{1}{\beta^\alpha}v\right)$$

ce qui montre que les variables $v_j = |x_j|^\alpha$ sont distribuées selon une loi gamma de paramètres $(1/\alpha, 1/\beta^\alpha)$.

Par ailleurs, comme la somme S_n de n variables aléatoires mutuellement indépendantes et distribuées selon des loi gamma de paramètres (a_i, b) est également distribuée selon un loi gamma de paramètres $(\sum_{i=1}^n a_i, b)$ [35], il en résulte que la norme $\varepsilon = \sum_{j=1}^n v_j = \sum_{j=1}^n |x_j|^\alpha$ est distribuée selon un loi gamma de paramètres $(a = n/\alpha, b = 1/\beta^\alpha)$. ■

A partir de cela, nous pouvons déduire la proposition suivante

Proposition 4 *Sous les hypothèses suivantes :*

- (H1) les n éléments de chaque vecteur source dans une classes S_k sont mutuellement indépendants et identiquement distribués suivant une densité de gaussienne généralisée de paramètre α_k, β_k.

- (H2) les éléments des vecteurs appartenant aux différentes classes sont indépendant et identiquement distribués suivant des densités de gaussienne généralisée de paramètres de puissance identique $\alpha_k = \alpha$ mais avec des paramètres de forme β_k différents.

Les paramètres de la distribution conjointe des vecteurs de la source qui suit un modèle MMGG, sont déduits de la manière suivante

$$\alpha_k = \alpha, \textit{ et } \beta_k = (1/b_k)^{1/\alpha} \quad \forall k = 1, ..., N_{mix}, \tag{4.25}$$

Preuve. C'est une généralisation de la proposition précédente au cas du modèle de mélange. L'hypothèse (H1) permet simplement d'écrire que la norme de chaque vecteur appartenant à la classe S_k, noté $X^{(k)}$, est distribuée selon une loi gamma de paramètres $(a_k = n/\alpha_k, b_k = 1/(\beta_k)^{\alpha})$. De plus, l'hypothèse (H2) permet d'écrire que les différentes classes sont distinguées par leur paramètre de forme et permet de déduire à partir de la proposition précédente que les paramètres du mélange MMGG sont les suivants :

$$\begin{cases} \alpha_k = \dfrac{a_k}{n} = \alpha, \\ \beta_k = (1/b_k)^{1/\alpha_k} = (1/b_k)^{1/\alpha}, \end{cases}$$

■

Ainsi, la distribution conjointe de chaque vecteur X conditionnellement à l'état S_k est :

$$f\left(X | \|X\|_\alpha^\alpha \in S_k\right) = \prod_{j=1}^{n} \mathcal{GG}(x_j; \alpha_k, \beta_k). \tag{4.26}$$

Avant de montrer dans le paragraphe suivant l'efficacité du modèle MMGG en évaluant les modèles R-D correspondants pour la QVA, nous insistons sur quelques points importants :

1. Le modèle MMGG est différent d'un modèle consistant en un mélange de distributions gaussiennes généralisées scalaires avec des étiquettes indépendantes parce que celui-ci ne prend pas en compte les dépendances des échantillons.
2. Les dépendances entre échantillons sont estimées d'une manière simple puisque le modèle MMGG est déduit des paramètres du mélange de gamma avec des étiquettes indépendantes représentant la distribution de la norme des vecteurs.
3. Comme il sera montré par la suite, notre approche permet de réutiliser les travaux relatifs à l'hypothèse classique i.i.d pour bénéficier de propriétés mathématiques avantageuses. Les coordonnées des vecteurs sont supposées identiquement distribuées, les distributions conditionnelles sont séparables et, finalement, la densité gaussienne généralisée i.i.d est une hypothèse très courante en QVA. En d'autres termes, la densité gaussienne généralisée i.i.d peut être vue comme un cas particulier de notre modèle MMGG.

4.3.3 Modèle débit-distorsion pour la QVA

Notre but est de concevoir des modèles débit-distorsion qui ne soient pas seulement fonction de la cellule du réseau mais également de la distribution de la source pour approximer correctement la fonction débit-distorsion à n'importe quel débit. Raffy *et al* [89] ont proposé de tels modèles. Ils ont d'abord généralisé le modèle de distorsion de Jeong et Gibson [53] sur des quantificateurs cubiques à débit fixé pour les réseaux réguliers de points à partir de l'hypothèse haute résolution. Ensuite, ils ont donné une estimation qui est valide pour une distribution i.i.d de type gaussienne généralisée. Dans ce paragraphe, nous généralisons leurs modèles débit-distorsion non-asymptotiques valables pour une distribution i.i.d de type gaussienne généralisée.

4.3.3.1 Modèle de débit-distorsion basé sur une distribution MMGG

Le point clef de la conception d'un modèle débit-distorsion basé sur une distribution MMGG est qu'il peut être calculé en transposant les modèles débit-distorsion dans le cas i.i.d. Ainsi, à partir

de (4.23), le modèle de distorsion D_{mix} associé à un mélange de densité est donné par :

$$D_{mix} = \int_{\mathbb{R}^n} \left\| X - \tilde{X} \right\|_2^2 f(X)\, dx_1 ... dx_n \tag{4.27}$$

$$= \sum_{k=1}^{N_{mix}} c_k \left[\int_{\mathbb{R}^n} \left\| X - \tilde{X} \right\|_2^2 f\left(X | \, \|X\|_\alpha^\alpha \in S_k\right)\, dx_1 ... dx_n \right], \tag{4.28}$$

où $\tilde{X} = \gamma Q(X/\gamma)$ est le vecteur reconstruit après une QVA. Finalement, à partir de (4.26), nous obtenons

$$D_{mix} = \sum_{k=1}^{N_{mix}} c_k \left[\int_{\mathbb{R}^n} \left\| X - \tilde{X} \right\|_2^2 \prod_{j=1}^{n} \mathcal{GG}(x_j; \alpha_k, \beta_k) dx_1 ... dx_n \right], \tag{4.29}$$

$$= \sum_{k=1}^{N_{mix}} c_k D_k \tag{4.30}$$

où D_k est la distorsion de la $k^{iéme}$ distribution du mélange. D_k est calculée pratiquement en utilisant les résultats de [86] et [89] qui donnent son expression analytique pour des sources i.i.d gaussienne et laplacienne.

Ainsi, si nous avons une source centrée de distribution gaussienne d'écart type σ_k, nous avons :

$$D_k(\gamma, \sigma_k) = \frac{\gamma^2}{12} \left[1 + \frac{12}{\pi^2} \sum_{j=1}^{+\infty} \frac{(1)^j}{j^2} e^{\frac{2\sigma_k^2 \pi^2 j^2}{\gamma^2}} \right] \tag{4.31}$$

De la même manière, pour une source centrée de distributon laplacienne d'écart type σ_k, on a :

$$D_k(\gamma, \sigma_k) = \frac{\gamma^2}{12} \left[1 + \frac{12}{\pi^2} \sum_{j=1}^{+\infty} \frac{(1)^j}{j^2} \frac{\gamma^2}{\gamma^2 + 2\sigma_j^2 \pi^2} \right] \tag{4.32}$$

Pour résumer, en remplaçant (4.31) ou (4.32) dans (4.30), nous obtenons une formule analytique de D_{mix}.

La formule (2.15) du chapitre 2 donne le débit pour un codage par QVA. En accord avec cette formule, le modèle de débit dans le cas d'une distribution MMGG peut s'écrire :

$$R_{mix} = -\sum_{r=0}^{r_T} P_{mix}(e = r) \left\{ \log_2(P_{mix}(e = r)) - \lceil \log_2(N(r)) \rceil \right\} \tag{4.33}$$

où $P_{mix}(e = r) = \sum_{k=1}^{N_{mix}} c_k P_k(e = r)$ est la probabilité discrète du rayon pour une distribution de la source de type MMGG et P_k est la probabilité de la $k^{iéme}$ classe S_k.

De la même manière que pour le calcul de D_k, une solution analytique de la loi du rayon $P_k(e = r)$ dans le cas s'une source de distribution laplacienne d'écart type σ_k a été proposée dans [79]. Elle peut s'écrire de la manière suivante

$$P_k(e = r) = \sum_{n_0=0}^{n} Card(r, n_0) \times \left(1 - \exp\left(-\frac{\lambda\gamma}{2}\right)\right)^{n_0} \times \left(\sinh\left(\frac{\lambda\gamma}{2}\right)\right)^{n-n_0} \times \exp(-\lambda\gamma r) \tag{4.34}$$

avec

$$\lambda = \frac{\sqrt{2}}{\sigma_k} \text{ et } Card(r, n_0) = cardinal\left(\left\{ \begin{array}{c} Y = \{y_1, ..., y_n\} \in \mathbb{Z}^n \mid \sum_{i=1}^{n} |y_i| = r \\ \text{et } cardinal\left(\{y_i \mid y_i = 0\}\right) = n_0 \end{array} \right\}\right) \tag{4.35}$$

où n_0 correspond au nombre de composantes nulles du vecteur Y quantifié. Nous rappelons que dans le cas laplacien ($\alpha = 1$), nous avons $e = \|Y\|$ (chapitre 2)

L'énumération du nombre de vecteurs $Card(r, n_0)$ sur chaque rayon r du réseau $\mathbb{Z}^n$ ayant n_0 composantes nulles, formule (4.35), peut être calculée en utilisant une partie des méthodes d'indexage par code préfixe [75]. *Voinson et al* [110] ont ainsi établi la formule analytique suivante :

$$Card(r, n_0) = 2^{n-n_0} \frac{n!}{n_0!} \sum_{y_{i_1}=1}^{\left\lfloor \frac{r}{\overline{n_0}} \right\rfloor} \sum_{y_{i_2}=y_{i_1}}^{\left\lfloor \frac{r-y_{i_1}}{\overline{n_0}-1} \right\rfloor} ... \sum_{y_{i_{\overline{n_0}}}=y_{i_{(\overline{n_0}-1)}}}^{\left\lfloor r-\sum_{j=1}^{\overline{n_0}-1} y_{i_j} \right\rfloor} \left(\prod_{y_{i_k}=y_{i_1}}^{\left\lfloor r-\sum_{j=1}^{\overline{n_0}-1} y_{i_j} \right\rfloor} \frac{}{\boldsymbol{\delta}\left(y_{i_k} - y_{i_{(k-1)}}\right) \times w\left(y_{i_k}\right)!} \right)^{-1} \tag{4.36}$$

$$\text{tel que} \quad \sum_{j=1}^{\overline{n_0}} y_{i_j} = r \;\; et \;\; w\left(y_{i_k}\right) = \sum_{q=y_{i_1}}^{y_{i_{\overline{n_0}}}} \boldsymbol{\delta}\left(y_{i_k} - q\right)$$

où $\boldsymbol{\delta}(u)$ est la fonction de Kronecker [5], $\overline{n_0}$ correspond au nombre de composantes non nulles du vecteur Y et y_{i_k} est la $i^{ème}$ composante non nulle.

Intéressons-nous maintenant aux résultats de l'estimation des courbes débit-distorsion à partir de ce modèle MMGG.

4.3.3.2 Résultats expérimentaux

Les expériences ont été réalisées en utilisant une QVA pyramidale qui comme on l'a déjà évoqué est un bon compromis entre les performances de codage et la complexité. Nous évaluons les performances du modèle de mélange proposé sur toutes les sous-bandes d'une TO2D sur cinq niveaux (filtre 9.7 de Daubechies) de plusieurs images naturelles (Lena, boat et peppers). En effet, dans l'optique d'une allocation, il est important de vérifier la validité du modèle sur toutes les sous-bandes. Nous avons également calculé les courbes débit-distorsion à partir d'un modèle classique (i.i.d laplacien) [89]. Elles sont évaluées en mettant la variance du modèle égale à celle de la sous-bande mesurée. Ces deux modèles sont ensuite comparés aux courbes débit-distorsion expérimentales obtenues à partir des données réelles. Pour faciliter la comparaison sur la précision des modèles, nous définissons l'erreur relative moyenne (ER_{moy}) et l'erreur relative maximale ($ER_{\max}$) - entre un modèle débit-distorsion de distorsion D et la courbe débit-distorsion de distorsion $D_{réel}$ en utilisant N_{test} valeurs de test :

$$ER_{moy} = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} \frac{|D(i) - D_{réel}(i)|}{D_{réel}(i)} \tag{4.37}$$

$$ER_{\max} = \max_{i=1,...,N_{test}} \frac{|D(i) - D_{réel}(i)|}{D_{réel}(i)} \tag{4.38}$$

Le tableau 4.1 permet de comparer les deux erreurs introduites dans les équations sur la sous-bande verticale de niveau 2 des trois images testées (Lena, boat et peppers). Comme nous pouvons

[5] $\boldsymbol{\delta}(u) = \left\{ \begin{array}{ll} 1 & \text{si } u = 0 \\ 0 & \text{sinon} \end{array} \right.$ Remarque : $\overline{\boldsymbol{\delta}(u)} = \left\{ \begin{array}{ll} 0 & \text{si } u = 0 \\ 1 & \text{sinon} \end{array} \right.$

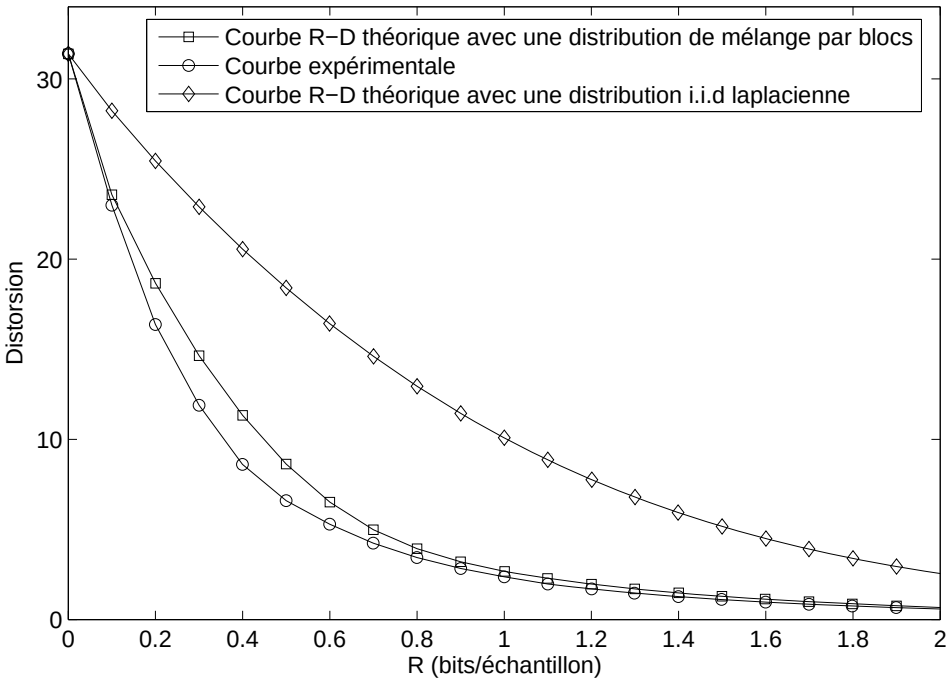

FIG. 4.8: Fonctions débit-distorsion pour la QVA - sous-bande V2 de l'image de Lena : données réelles, modèle de mélange proposé, modèle i.i.d laplacien - vecteurs de taille (4×2).

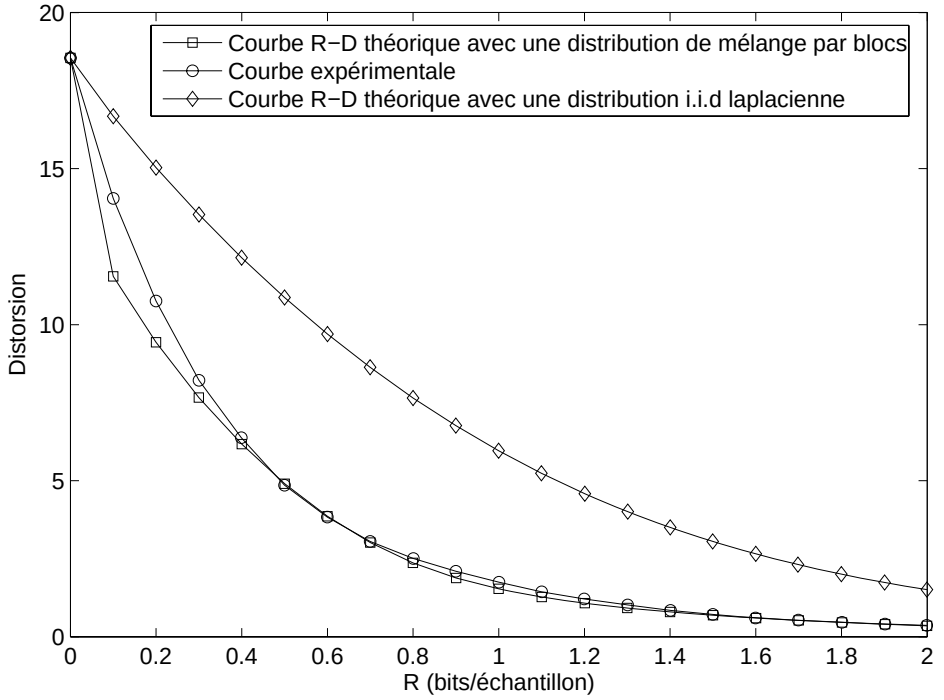

FIG. 4.9: Fonctions débit-distorsion pour la QVA - sous-bande H3 de l'image de Lena : données réelles, modèle de mélange proposé, modèle i.i.d laplacien - vecteurs de taille (2×4).

constater, l'hypothèse du mélange donne un modèle bien plus réaliste sur cette sous-bande. Cette efficacité est confirmée sur les figures 4.8, 4.9 et 4.10 où le modèle de mélange correspond très bien aux courbes R-D des données réelles tandis que la courbe générée à partir de l'hypothèse i.i.d diverge. Le tableau 4.3 montre que le modèle de mélange par blocs est supérieur pour toutes les sous-bandes de Lena, à l'exception de D1. Cela s'explique par le fait que cette sous-bande suit une loi laplacienne i.i.d. Notons tout de même que N_{mix} est un paramètre réglable et ainsi on peut le mettre aisément N_{mix} à 1. Soulignons enfin que le modèle i.i.d a une précision plus proche de notre modèle de mélange R-D dans les niveaux correspondant aux faibles résolutions (typiquement 5). Ce point peut s'expliquer par le fait que les coefficients d'ondelettes sont ici clairsemés et les "clusters" sont plus grands que dans les résolutions fines (qui correspondent aux contours et textures de l'image). Par ailleurs, la modélisation de ces sous-bandes de petite taille est moins importante en terme de complexité car le calcul de leur courbe expérimentale est négligeable. Il est tout à fait possible d'envisager une procédure d'allocation à faible complexité calculatoire où les courbes des petites sous-bandes sont calculées à partir des données réelles et les courbes correspondants aux autres sous-bandes sont évaluées à partir du modèle MMGG.

	Lena	Boat	Peppers
ER_{moy}	**0,16** ; 2,57	**0,30** ; 1,78	**0,07** ; 2,96
$ER_{\max}$	**0,37** ; 3,64	**0,52** ; 3,49	**0,23** ; 4,07

Tab. 4.1: Erreur relative - moyenne (ER_{moy}) et maximale ($ER_{\max}$) entre les fonctions R-D expérimentales par QVA et les modèles analytiques (i.i.d laplacien et modèle de mélange proposé en gras). Les tests ont été réalisés sur la sous-bande V2 des images, Lena, boat et peppers.

	Lena	Boat	Peppers
ER_{moy}	**0.04** ; 2.87	**0.20** ; 0.44	**0.08** ; 3.33
$ER_{\max}$	**0.08** ; 4.1	**0.32** ; 2.66	**0.17** ; 4.43

Tab. 4.2: Erreur relative - moyenne (ER_{moy}) et maximale ($ER_{\max}$) entre les fonctions R-D expérimentales par QVAZM et les modèles analytiques (i.i.d laplacien et modèle de mélange proposé en gras). Les tests ont été réalisés sur la sous-bande V2 des images, Lena, boat et peppers.

4.3.4 Modèle débit-distorsion pour la QVAZM

Les résultats expérimentaux précédents soulignent que la QVA utilisée dans le cadre classique i.i.d peut être améliorée en prenant en compte les dépendances statistiques à l'intérieur des vecteurs. Comme il a déjà été mentionné, l'une des principales propriétés d'un signal distribué par un mélange de densités par blocs est le grand nombre de vecteurs de faible énergie. Ce point suggère qu'un dictionnaire avec une forme adaptée peut surpasser des méthodes existantes, en exploitant mieux le gain de parcimonie. C'est dans ce cadre que la QVAZM a été proposée. Pour rappel, elle permet de diminuer le débit entropique de l'énergie en remplaçant par 0 les vecteurs dont l'énergie est en deçà d'un seuil correspondant au rayon de la zone morte permettant ainsi une quantification plus fine de la région uniforme. Nous proposons de transposer ici à la QVAZM notre modèle MMGG proposé dans le paragraphe précédent pour déduire un modèle débit-distorsion : point crucial dans la réalisation pratique de la QVAZM. En effet, comme nous l'avons déjà évoqué, l'allocation de débits de notre méthode requiert le réglage d'un paramètre supplémentaire : la zone morte vectorielle.

De manière analogue à la modélisation débit-distorsion pour la QVA classique grâce à un modèle MMGG, la conception d'un modèle débit-distorsion pour la QVAZM à partir d'une distribution MMGG nécessite uniquement de calculer les modèles débit-distorsion sous l'hypothèse gaussienne généralisée.

4.3.4.1 Modèle de distorsion

Le modèle de distorsion de la QVAZM consiste à calculer la distorsion dans les trois zones de la figure 2.12 du chapitre 2, à savoir la zone morte vectorielle (ZM), la zone de surcharge (ZS) et la zone uniforme (ZU). Nous approximons la distribution du bruit de quantification dans la zone de surcharge et la zone uniforme par une loi uniforme. Dans la suite nous considérons que la zone de surcharge est incluse dans la zone uniforme. Le modèle de distorsion est déduit de la proposition suivante :

Proposition 5 *Sous l'hypothèse haute-résolution dans la région uniforme, la distorsion en fonction du rayon de la zone morte R_{ZM} et du facteur de projection γ est donnée par :*

$$D\left(R_{ZM},\gamma\right)=\frac{1}{n}\left[D_{ZM}\left(R_{ZM}\right)+D_{ZU}(\gamma)\left(1-F_n\left(R_{ZM}\right)\right)\right] \tag{4.39}$$

où $D_{ZM}(R_{ZM})$ est la distorsion à l'intérieur de la zone morte vectorielle , $D_{ZU}(\gamma)$ est la distorsion dans la zone uniforme ZU, F_n est la fonction de répartition de la loi de $\|X\|_\alpha^\alpha$ et n la dimension de X. $D_{ZU}(\gamma)$ correspond à la distorsion granulaire d'un réseau Λ pour une source uniformément distribuée.

Dans le cas d'un réseau $\mathbb{Z}^n$, nous avons $D_{ZU}(\gamma)=\frac{n\gamma^2}{12}$. De plus pour une source gaussienne généralisée i.i.d, la formule $D_{ZM}(R_{ZM})$ est donnée par [6] :

$$D_{ZM}\left(R_{ZM}\right)=n\int\limits_{|x|^\alpha<R_{ZM}} x^2\mathcal{GG}(x)F_{n-1}\left(R_{ZM}-|x|^\alpha\right)dx \tag{4.40}$$

Preuve. L'expression analytique de D_{ZM} est calculée en utilisant des propriétés particulières de la densité de probabilité conjointe d'une variable aléatoire distribuée suivant une loi i.i.d (4.26). Nous avons :

$$D_{ZM}\left(R_{ZM}\right)=\int_{ZM}\left(x_1^2+...+x_n^2\right)\left(\prod_{k=1}^{n}\mathcal{GG}(x_k)dx_k\right)$$

où $ZM=\{X;\|X\|_\alpha^\alpha<R_{ZM}\}$. Ainsi,

$$D_{ZM}\left(R_{ZM}\right)=\sum_{i=1}^{n}\int_{ZM}x_i^2\left(\prod_{k=1}^{n}\mathcal{GG}(x_k)dx_k\right)$$

Puisque nous avons aussi :

$$\forall i,j;\int_{ZM}x_i^2\left(\prod_{k=1}^{n}\mathcal{GG}(x_k)dx_k\right)=\int_{ZM}x_j^2\left(\prod_{k=1}^{n}\mathcal{GG}(x_k)dx_k\right)$$

Nous obtenons :

$$D_{ZM}\left(R_{ZM}\right)=n\int_{ZM}x_1^2\left(\prod_{k=1}^{n}\mathcal{GG}(x_k)dx_k\right)$$

Nous notons $\mathbb{A}=\{X;|x_2|^\alpha+...+|x_n|^\alpha<R_{ZM}-|x_1|^\alpha\}$
Nous avons :

$$\{\|X\|_\alpha^\alpha<R_{ZM}\}=\{|x_1|<R_{ZM}\}\cap\mathbb{A}$$

[6] Pour des soucis de simplicité $\mathcal{GG}(x;\alpha,\beta)$ est remplacé ici par $\mathcal{GG}(x)$.

Ainsi,

$$D_{ZM}\left(R_{ZM}\right) = n \int\limits_{|x|^\alpha < R_{ZM}} x^2 \mathcal{GG}(x) \left[\int\limits_{\mathbb{A}} \left(\prod_{k=2}^{n} \mathcal{GG}(x_k) dx_k \right) \right] dx$$

L'intégrale entre crochet correspond à la fonction de répartition F_{n-1} de la loi du rayon pour une distribution gaussienne généralisée de dimension $n-1$. Nous obtenons donc

$$D_{ZM}\left(R_{ZM}\right) = n \int\limits_{|x|^\alpha < R_{ZM}} x^2 \mathcal{GG}(x) F_{n-1}\left(R_{ZM} - |x|^\alpha\right) dx$$

■

Ainsi D_{ZM} dépend seulement de la connaissance de la fonction de répartition F_n de la loi $\|X\|_\alpha^\alpha$. Pour une distribution de type gaussienne généralisée, l'expression analytique de F_n est donnée dans [18]. Dans le cas laplacien, la fonction de répartition du rayon F_n est obtenue par une intégration par parties successives et peut s'écrire

$$F_n\left(R\right) = 1 - e^{-\lambda R} \sum_{j=0}^{n} \frac{(\lambda R)^j}{j!} \tag{4.41}$$

avec $\lambda = \frac{\sqrt{2}}{\sigma}$

Comme pour la QVA, le modèle débit-distorsion de mélange sera testé ici sous l'hypothèse d'un mélange de densités laplaciennes multi-dimensionnelles. La distorsion dans la zone morte dans le cas d'une source laplacienne i.i.d d'écart type σ est donnée par :

$$D_{ZM}\left(R_{ZM}\right) = n \left(J - 2\lambda e^{-\lambda R_{ZM}} R_{ZM}^3 \sum_{j=0}^{2} \frac{(\lambda R_{ZM})^j}{(j+3)!} \right) \tag{4.42}$$

avec $J = -e^{-\lambda R_{ZM}} \left(R_{ZM}^2 + \frac{2}{\lambda} R_{ZM} + \frac{2}{\lambda^2} \right) + \frac{2}{\lambda^2}$.

Cette expression analytique est obtenue en injectant la fonction de répartition définie en (4.41) dans (4.40).

Finalement, le modèle de mélange débit-distorsion consiste simplement à substituer la formule (4.39) dans la formule suivante (correspondant à l'équation 4.30) :

$$D_{mix} = \sum_{k=1}^{N_{mix}} c_k D_k = \sum_{k=1}^{N_{mix}} c_k D_k(R_{ZM_k}, \gamma_k, \sigma_k) \tag{4.43}$$

Une fois le modèle de distorsion de la QVAZM calculé pour une distribution MMGG, intéressons-nous au modèle de débit pour la QVAZM.

4.3.4.2 Modèle de débit

Le modèle de débit est déduit de la loi du rayon classique en remplaçant dans la formule suivante (correspondant à l'équation 4.33) :

$$R_{mix} = -\sum_{r=0}^{r_T} P_{mix}\left(e = r\right) \left\{ \log_2(P_{mix}\left(e = r\right)) - \lceil \log_2(N\left(r\right)) \rceil \right\}$$

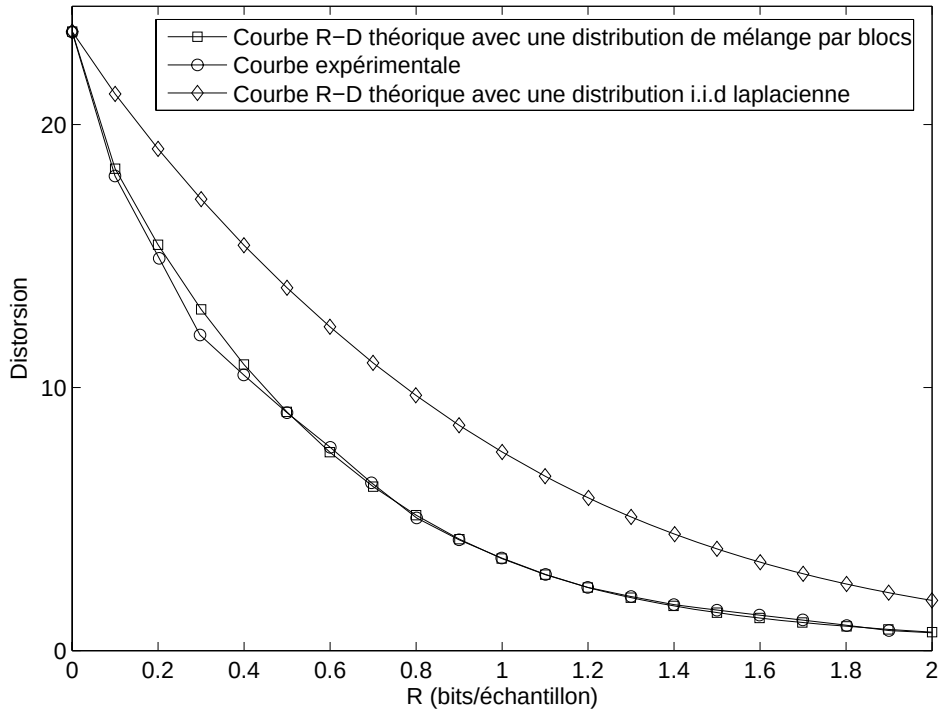

FIG. 4.10: Fonctions débit-distorsion pour la QVA - sous-bande D4 de Lena : données réelles, modèle de mélange proposé, modèle i.i.d laplacien - vecteurs de taille (2×2).

la probabilité du vecteur nul par la probabilité d'appartenir à la zone morte vectorielle de rayon R_{ZM}. La probabilité discrète du rayon P_{ZM} peut s'écrire :

$$\begin{aligned}
P_{ZM}(e=0) &= \sum_{k=1}^{N_{mix}} c_k P_k\left(\|X\|_\alpha^\alpha < R_{ZM}\right) \\
P_{ZM}(e=1) &= ... = P_{ZM}(e=\delta-1) = 0 \\
P_{ZM}(e=\delta) &= \left[\sum_{r=0}^{\delta} P_{mix}(e=r)\right] - P_{ZM}(e=0) \\
P_{ZM}(e=r) &= P_{mix}(e=r) = \sum_{k=1}^{N_{mix}} c_k P_k(e=r),\ r>\delta
\end{aligned}$$

où P_k correspond à la probabilité relative au $k^{ème}$ état de la source (ou de la norme). On rappelle que $\delta = \left\lfloor \frac{R_{ZM}}{\gamma} \right\rfloor + 1$ est le rayon de la première couche hors zone morte.

Le modèle de débit dans le cas d'une QVAZM laplacienne est donné dans [110]. Sous cette hypothèse laplacienne, les probabilité discrètes $P_{ZM}(e=r)$ pour $r>\delta$, servant au calcul du débit dans la ZU sont calculées à partir de l'équation (4.34). La probabilité $P_{ZM}(e=0)$ relative au débit de la zone morte est obtenue à partir de la fonction de répartition du rayon définie en (4.41)

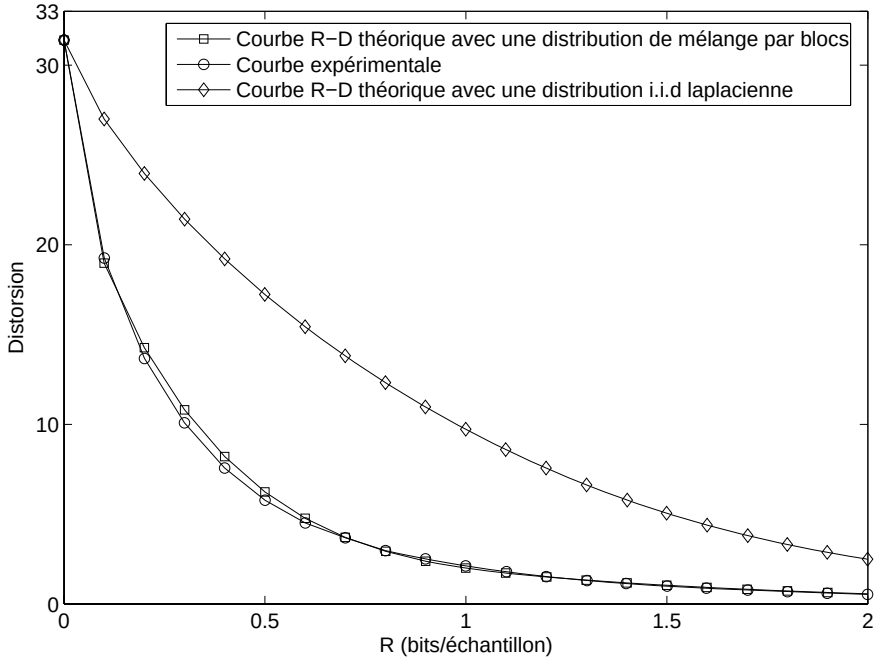

Fig. 4.11: Fonctions débit-distorsion pour la QVAZM - sous-bande V2 de l'image de Lena : données réelles, modèle de mélange proposé, modèle i.i.d laplacien - vecteurs de taille (4×2).

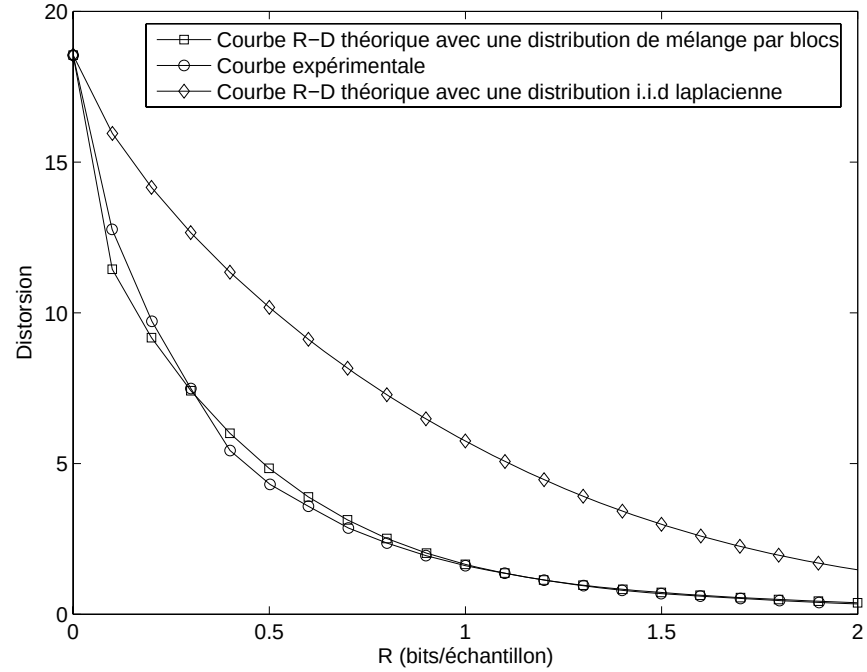

Fig. 4.12: Fonctions débit-distorsion pour la QVAZM - sous-bande H3 de l'image de Lena : données réelles, modèle de mélange proposé, modèle i.i.d laplacien - vecteurs de taille (2×4).

		QVA		QVAZM	
Sous bandes	Mesure	MMGG	i.i.d classique	MMGG	i.i.d classique
V1	ER_{moy}	**0,07**	1,19	**0,08**	1,36
	$ER_{\max}$	**0,18**	1,53	**0,19**	1,79
H1	ER_{moy}	**0,06**	0,39	**0,05**	0,48
	$ER_{\max}$	**0,18**	0,53	**0,11**	0,65
D1	ER_{moy}	0,07	**0,05**	**0,07**	0,08
	$ER_{\max}$	**0,12**	0,10	**0,13**	0,15
V2	ER_{moy}	**0,16**	2,57	**0,04**	2,87
	$ER_{\max}$	**0,32**	3,67	**0,08**	4,1
H2	ER_{moy}	**0,07**	2,19	**0,07**	2,4
	$ER_{\max}$	**0,18**	2,94	**0,12**	3,24
D2	ER_{moy}	**0,06**	1,57	**0,05**	1,8
	$ER_{\max}$	**0,11**	2,17	**0,09**	2,4
V3	ER_{moy}	**0,06**	1,29	**0,05**	1,4
	$ER_{\max}$	**0,15**	2,35	**0,13**	2,43
H3	ER_{moy}	**0,05**	2,1	**0,06**	2,2
	$ER_{\max}$	**0,18**	3,4	**0,13**	3,3
D3	ER_{moy}	**0,07**	2,53	**0,16**	2,6
	$ER_{\max}$	**0,16**	4,11	**0,25**	4,2
V4	ER_{moy}	**0,1**	0,24	**0,04**	0,20
	$ER_{\max}$	**0,2**	0,61	**0,09**	0,46
H4	ER_{moy}	**0,18**	0,8	**0,11**	0,7
	$ER_{\max}$	**0,38**	1.42	**0.17**	1.28
D4	ER_{moy}	**0.03**	1.04	**0.07**	0.94
	$ER_{\max}$	**0.08**	1.88	**0.15**	1.65
V5	ER_{moy}	**0.11**	0.17	0.18	**0.03**
	$ER_{\max}$	**0.29**	0.30	0.36	**0.1**
H5	ER_{moy}	0.06	**0.035**	**0.16**	0.25
	$ER_{\max}$	**0.19**	0.58	**0.3**	0.5
D5	ER_{moy}	**0.25**	0.38	**0.06**	0.30
	$ER_{\max}$	**0.31**	0.76	**0.14**	0.68

TAB. 4.3: Erreur relative - moyenne (ER_{moy}) et maximale (ER_{max}) entre les fonctions R-D expérimentales par QVA / QVAZM et les modèles analytiques (i.i.d laplacien et modèle de mélange proposé en gras). Les tests ont été réalisés sur toutes les sous-bandes de l'image de Lena.

4.3.4.3 Résultats expérimentaux

Nous avons réalisé les mêmes tests que pour la QVA sur toutes les sous-bandes des trois images naturelles évaluées (tableau 4.2). Les figures 4.11 et 4.12 montrent la très bonne précision de notre modèle analytique pour estimer les fonctions débit-distorsion de la QVAZM. De plus, le tableau 4.3 qui compare la précision du modèle de mélange pour la QVA et la QVAZM permet de constater que le modèle de mélange pour la QVAZM est un tout petit plus précis que celui de la QVA pour la plupart des sous-bandes. Enfin dans le but d'utiliser ces modèles pour la procédure d'allocation de la QVAZM 3D, nous l'avons testé sur les sous-bandes produites par une TO3D sur une image médicale volumique. La figure 4.13 montre la précision du modèle sur la sous-bande LHL1 d'une TO3D sur 4 niveaux de l'image B de la base E3.

4.3.4.4 Complexité

Le coût calculatoire de notre approche statistique est faible dans l'optique d'une allocation de débits. Pour une sous-bande, les paramètres du modèle de mélange par blocs sont estimés par un algorithme MCMC qui converge presque instantanément. Pratiquement, nous utilisons pour cette partie une fonction écrite en Matlab [76]. A titre d'exemple, bien que non optimisée, cette fonction calcule les paramètres du mélange pour les 15 sous-bandes (TO2D sur 5 niveaux) de l'image Lena en environ 2 secondes[7]. Cette partie pourra être améliorée en utilisant des algorithmes plus rapides pour l'estimation des paramètres du modèle MMGG. Néanmoins, c'est la seule partie de l'allocation où nous faisons appel aux données réelles. Une fois les paramètres du modèle MMGG établis, les distorsions et les débits sont uniquement évalués à partir des paramètres du modèle. Ainsi, la minimisation sous contrainte d'égalité résolue par une approche lagrangienne devient très rapide. Cette réduction de complexité est d'autant plus importante pour des sous-bandes d'une pile d'images médicales produites par une TO3D. En effet, la taille des données des plus grandes sous-bandes peut être jusqu'à 250 fois plus grande que celle des plus grandes sous-bandes des images 2D. On comprend aisément l'importance d'un modèle statistique pour les images médicales 3D.

Etant donné, ce modèle analytique de mélange par blocs validé, nous pouvons désormais l'appliquer dans une procédure rapide d'allocation de débits pour la QVAZM. Les résultats de celle-ci seront comparés à ceux d'une QVAZM par une allocation classique lagrangienne et par l'allocation analytique déduite de l'approximation exponentielle présentée dans le paragraphe 2.

4.4 Comparaisons des allocations de débits

Nous comparons ici les trois allocations proposées dans ce manuscrit pour la QVAZM. La première présentée dans le chapitre 2 utilise une approche classique lagrangienne pour résoudre le problème de minimisation. Les deux autres approches proposent des simulateurs pour calculer les fonctions $D_i(R_i)$. La deuxième approche proposée utilise un ajustement des courbes $D_i(R_i)$ en les approximant par un modèle exponentiel. Elle conduit à une simple résolution analytique du problème. La dernière méthode calcule les courbes débit-distorsion sans appel aux données à travers un modèle statistique MMGG. Ces courbes sont injectées dans le même algorithme d'optimisation que l'approche lagrangienne classique.

La méthode QVAZM a déjà démontré son efficacité sur les images naturelles : les résultats obtenus dans [110] et [111] classent la méthode au rang des standards que sont SPIHT et JPEG2000 en terme de compromis débit/distorsion. Par exemple, pour un taux de compression de 64 : 1, les PSNR obtenus sur l'image Lena sont de 30,90 dB pour la QVAZM avec un codage stack-run pour les normes ; 30,82 dB pour JPEG2000 et 31,08 dB pour SPIHT. En outre, elle a également prouvé son aptitude à préserver les structures des images [111]. De plus, on constate une amélioration

[7] Les tests ont été effectués sur un ordinateur portable avec processeur Intel T2300 1,66 Ghz.

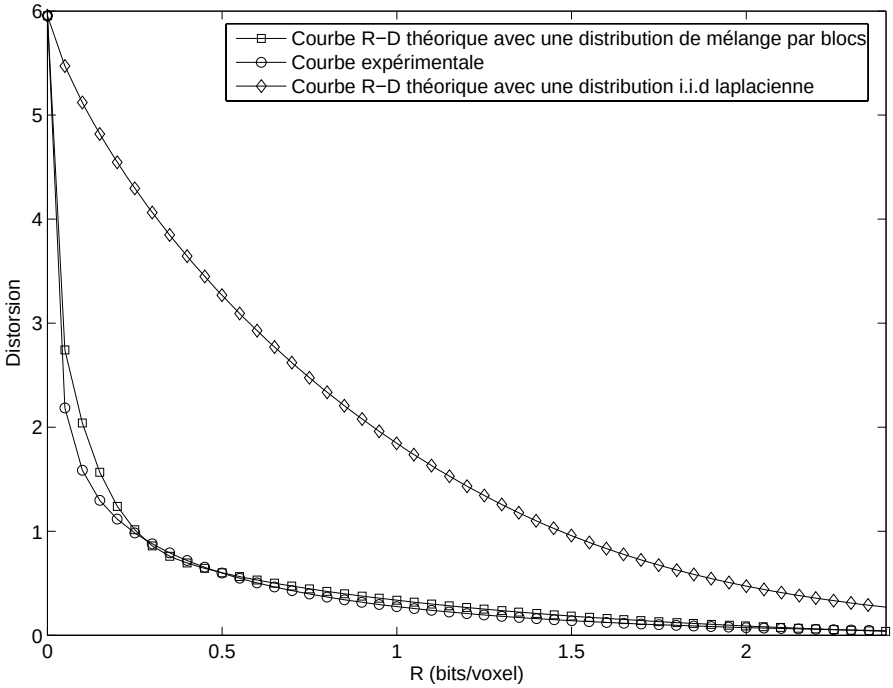

FIG. 4.13: Fonctions débit-distorsion pour la QVAZM 3D - sous-bande LHL1 de la pile d'images B de la base E3 : données réelles, modèle de mélange proposé, modèle i.i.d laplacien - vecteurs de taille (2×2×2).

significative de la qualité visuelle de la QVAZM à bas débit, en particulier sur les zones texturées. Par exemple : si nous observons le chapeau de l'image *Lena* pour un TC = 64 : 1 (figure 4.15) ou pour un TC = 103 : 1 (figure 4.14) , nous pouvons constater que certains détails sont mieux préservés avec notre méthode que celle de SPIHT ou de JPEG2000. Nous pouvons également observer l'amélioration de la qualité visuelle sur le zoom effectué sur les yeux de *Lena* pour un TC = 64 : 1 (figure 4.16), en particulier sur l'oeil gauche. Les différentes images que nous présentons à la fin du chapitre permettent d'apprécier cette qualité visuelle.

Nous avons déjà évoqué la réduction notoire de la complexité des allocation de débits proposées par rapport à une approche lagrangienne classique. Les tests effectués ont pour but de valider les deux allocations et de comparer leur résultat à une allocation de type lagrangienne. Le tableau 4.4 permet d'effectuer des comparaisons en terme de PSNR entre les trois approches pour trois débits cible : 0,125 bit/pixel (taux de compression de 1: 64), 0,25 bit/pixel (taux de compression de 1: 32) et 0,5 bits/pixel (taux de compression de 1: 16). Les résultats montrent que les deux modèles proposées présentent des résultats très proches de l'allocation lagrangienne pure. Nous notons toutefois une plus grande précision de l'approche statistique par mélange de blocs qui est toujours légèrement supérieure au modèle exponentiel. En effet, les courbes débit-distorsion du modèle MMGG sont construites à partir d'une modélisation bien adaptée à la statistique de la source, au contraire de celles du modèle exponentiel qui suivent une fonction moins flexible de type $D(R) = Ce^{-aR}$. Ainsi, la forme des courbes de certaines sous-bandes ne peut être totalement bien modélisée à travers cette fonction. De plus, ce même tableau 4.4 permet de comparer les performances de notre algorithme avec les différentes allocations proposées à celles de l'algorithme SPIHT et du standard JPEG2000. En ce qui concerne les comparaisons visuelles, la différence en terme de qualité visuelle est négligeable entre les trois allocations comme le montre les images à la

Débit	Méthode (allocation)	Lena	boat	peppers
0,125 bit/pixel	QVAZM (lagrangienne)	31,02 dB	27,98 dB	30,30 dB
	QVAZM (exponentielle)	30,96 dB	27,90 dB	30,03 dB
	QVAZM (MMGG)	30,98 dB	27,93 dB	30,30 dB
	SPIHT	31,08 dB	28,16 dB	30,65 dB
	JPEG 2000	30,90 dB	27,99 dB	30,56 dB
0,25 bit/pixel	QVAZM (lagrangienne)	34,03 dB	30,81 dB	33,14 dB
	QVAZM (exponentielle)	33,80 dB	30,64 dB	32,54 dB
	QVAZM (MMGG)	33,97 dB	30,70 dB	33,13 dB
	SPIHT	34,11 dB	30,97 dB	33,47 dB
	JPEG2000	34,02 dB	30,94 dB	33,41 dB
0,5 bit/pixel	QVAZM (lagrangienne)	37,19 dB	34,20 dB	35,61 dB
	QVAZM (exponentielle)	36,81 dB	34,02 dB	35.20 dB
	QVAZM (MMGG)	37,11 dB	34,04 dB	35,59 dB
	SPIHT	37,21 dB	34,45 dB	35,92 dB
	JPEG2000	37,19 dB	34,57 dB	35,79 dB

Tab. 4.4: Comparaison à 3 débits (0,125 bit/pixel, 025 bit/pixel et 0.5 bit/pixel) des 3 allocations proposées pour la QVAZM avec la méthode : lagrangienne, par modélisation exponentielle et par mélange de blocs. Les images testées sont Lena, boat et peppers.

fin du chapitre (figure 4.17 à 4.28).

Enfin, nous comparons l'allocation par modèle MMGG et par approche lagrangienne sur l'image médicale B de la base E3 testée dans le chapitre 2. Pour l'allocation basée sur le modèle MMGG, seules les sous-bandes des deux plus grands niveaux de résolution sont estimées à partir du modèle. Les plus petites sous-bandes dont la complexité est négligeable utilisent les données réelles pour une plus grande précision. Les résultats sont reportés dans le tableau 4.5 et montrent encore un fois que l'allocation s'appuyant sur le modèle MMGG est très proche de celle faisant appel aux données réelles.

Débit	0,375 bit/voxel		0,5 bit/voxel		0,75 bit/voxel	
Allocation	Lagrangienne	MMGG	Lagrangienne	MMGG	Lagrangienne	MMGG
PSNR	55,41 dB	55,25 dB	57,73 dB	57,58 dB	61,58 dB	61,53 dB

Tab. 4.5: PSNR moyen (dB) - Comparaison entre une allocation pour la QVAZM 3D avec la méthode : lagrangienne et par mélange de blocs à différents débits testés sur le scanner B de la base E3.

4.5 Conclusion

La mise en oeuvre d'un schéma de compression complet utilisant la QVAZM nécessite une procédure d'allocation de débits pour répartir au mieux les débits dans chaque sous-bande, minimisant ainsi la distorsion globale de l'image. Dans les approches intra-bandes comme la nôtre, cette étape d'allocation de débits peut s'avérer extrêmement coûteuse en terme de complexité calculatoire. Afin de s'affranchir de cette contrainte, nous avons proposé dans ce chapitre deux types d'allocations de débits rapides dédiées à notre méthode.

La première solution proposée utilise un ajustement des courbes débit-distorsion par un simple modèle exponentiel. Les paramètres de ce modèle sont estimés par une régression linéaire à partir de seulement trois points débit-distorsion calculés. Les courbes modélisées par cette fonction expo-

FIG. 4.14: Zoom sur le chapeau de Lena pour un TC = 103 : 1 (0,0777 bit/pixel), images de gauche à droite : originale, SPIHT, JPEG2000, QVAZM.

nentielle conduisent à une résolution analytique de l'allocation de débits négligeable pour le coût de l'optimisation total. Concrètement, l'allocation de débits de la QVAZM à partir du modèle exponentiel présente une complexité de 45 opérations/pixel à bas débits (jusqu'à 0,25 bits/pixel pour des images naturelles classiques comme boat ou Lena) tout en obtenant des résultats numériques et visuels très proches d'une allocation de type lagrangienne. Néanmoins, ce modèle souffre d'un manque de flexibilité et ne peut approximer les courbes R-D de certaines sources comme certaines sous-bandes d'une TO3D pour les images médicales.

Ce manque de flexibilité est résolu par la deuxième solution proposée qui appartient à une allocation de débits par approche statistique. Nous avons proposé des modèles débit-distorsion statistiques pour la QVA et la QVAZM permettant de prendre en compte les propriétés d'agglutinement et de parcimonie des coefficients d'ondelettes. Ils s'appuient sur une modélisation de la norme utilisant un mélange de lois gamma qui conduit à un mélange de densité de gaussiennes généralisées multidimensionnelles (MMGG) pour la distribution conjointe des vecteurs de la source eux-mêmes. Les modèles R-D sont simplement déduits du modèle classique dans le cas classiques i.i.d. Les résultats expérimentaux dans le cas d'une QVA/QVAZM pyramidale sous l'hypothèse d'un mélange de lois laplaciennes multidimensionnelles montre la précision de ce modèle. Ce dernier point confirme que la distribution MMGG est a priori bien adaptée pour des signaux comme les coefficients d'ondelettes.

L'efficacité du modèle MMGG pour l'estimation des courbes R-D de la QVAZM entraîne une allocation de débits de grande précision. Ainsi, les résultats numériques sont encore plus proches d'une QVAZM avec allocation lagrangienne classique. Encore une fois, les propriétés visuelles de la QVAZM (préservation des structures fines) sont conservées avec cette allocation. Par ailleurs, la flexibilité du modèle MMGG a permis de l'appliquer avec une grande précision sur des sous-bandes d'une d'un scanner ORL. Nous disposons ainsi d'un outil d'allocation de débits statistique rapide et efficace pour ce type d'images volumiques. Finalement, la précision du modèle MMGG et ses propriétés mathématiques avantageuses sont prometteuses pour envisager des applications dans d'autres domaines du codage de source ou pour d'autres problèmes du traitement du signal comme le débruitage.

Fig. 4.15: Zoom sur le chapeau de Lena pour un TC = 64 : 1 (0, 0125 bit/pixel), images de gauche à droite : originale, SPIHT, JPEG2000, QVAZM.

Fig. 4.16: Zoom sur les yeux de Lena pour un TC = 64 : 1 (0, 125 bit/pixel), images du haut : originale, QVAZM, images du bas : JPEG2000, SPIHT.

FIG. 4.17: Allocation des débits pour la QVAZM par le modèle de mélange de blocs - image Lena - débit : 0,125 bit/pixel ; PSNR = 30,98 dB.

FIG. 4.18: Allocation des débits pour la QVAZM par minimisation lagrangienne - image Lena - débit : 0,125 bit/pixel ; PSNR = 31,02 dB.

Fig. 4.19: Allocation des débits pour la QVAZM par le modèle exponentiel - image Lena - débit : 0,125 bit/pixel ; PSNR = 30,96 dB.

Fig. 4.20: Allocation des débits pour la QVAZM par le modèle de mélange de blocs - image Lena - débit : 0,5 bit/pixel ; PSNR = 37,11 dB.

FIG. 4.21: Allocation des débits pour la QVAZM par minimisation lagrangienne - image Lena - débit : 0,5 bit/pixel ; PSNR = 37,19 dB.

FIG. 4.22: Allocation des débits par le modèle exponentiel - image Lena - débit : 0,5 bit/pixel ; PSNR = 36,81 dB.

Fig. 4.23: Allocation des débits pour la QVAZM par le modèle de mélange de blocs - image peppers - débit : 0,125 bit/pixel ; PSNR = 30,30 dB.

Fig. 4.24: Allocation des débits pour la QVAZM par minimisation lagrangienne - image peppers - débit : 0,125 bit/pixel ; PSNR = 30,30 dB.

FIG. 4.25: Allocation des débits pour la QVAZM par le modèle exponentiel - image Lena - débit : 0,125 bit/pixel ; PSNR = 30,03 dB.

FIG. 4.26: Allocation des débits pour la QVAZM par le modèle de mélange de blocs - image boat - débit : 0,25 bit/pixel ; PSNR = 30,70 dB.

Fig. 4.27: Allocation des débits pour la QVAZM par minimisation lagrangienne - image boat - débit : 0,25 bit/pixel ; PSNR = 30,91 dB.

Fig. 4.28: Allocation des débits pour la QVAZM par le modèle exponentiel - image boat - débit : 0,25 bit/pixel ; PSNR = 30,64 dB.

Conclusion générale et perspectives

L'imagerie médicale ou plus exactement l'imagerie radiologique permet une investigation de plus en plus fine des organes humains. La contrepartie réside dans une masse de données générée chaque jour dans un service de radiologie considérable. La nécessité de compresser les images apparaît donc aujourd'hui incontournable pour remplir les fonctionnalités d'archivage et de transmission rapide. Dans ce manuscrit, nous avons exposé le fait que la compression dite "sans perte" ne permettait pas une réduction significative du volume de ces données. Nous avons ensuite investigué la compression "avec pertes" maîtrisées, à savoir des pertes n'affectant pas la qualité des images pour l'usage régulier par les praticiens.

Nous avons proposé une nouvelle méthode 3D basée sur une transformée en ondelettes 3D dyadique avec une étape de quantification utilisant une Quantification Vectorielle Algébrique avec Zone Morte 3D. La principale contribution de ce travail réside dans la conception d'une zone morte multidimensionnelle pendant l'étape de quantification qui permet de prendre en compte les corrélations entre les voxels voisins. La zone morte vectorielle a pour but de seuiller les vecteurs non significatifs (en terme d'énergie) afin de quantifier plus finement les vecteurs contenant les structures de l'image. Les résultats de notre méthode sont encourageants. Ils révèlent une supériorité numérique de notre méthode par rapport à plusieurs des meilleures méthodes actuelles (parmi lesquelles SPIHT 3D), sur différents scanners et IRM couramment utilisés pour évaluer la compression d'images médicales. En terme visuel, nous notons une supériorité de notre méthode qui préserve mieux les détails de ces images. Ce résultat est confirmé sur des scanners ORL à travers l'évaluation de deux spécialistes (un radiologue et un radiothérapeute) du domaine qui jugent notre méthode supérieure visuellement à l'algorithme SPIHT 3D. Par ailleurs, les scanners ORL comprimés jusqu'à un taux de compression de 16 : 1 sont jugés localement et globalement de très bonne qualité et apparaissent compatibles avec la pratique clinique classique pour les deux praticiens. Le radiothérapeute repousse lui cette limite d'acceptabilité de la compression jusqu'à un taux de compression de 24 : 1.

De plus, pour accréditer l'idée de la compression avec pertes maitrisées auprès de la communauté médicale, nous avons mesuré l'impact d'une compression avec pertes sur les performances d'un CAD (outil d'aide à la décision) performant pour la détection automatique de nodules pulmonaires et des embolies pulmonaires. Nous avons montré que la détection des nodules par le CAD n'était pas détériorée par la compression basée sur SPIHT 3D jusqu'à un niveau de compression de 48 : 1 et restait robuste jusqu'à 96 : 1. Une seconde étude sur la volumétrie a prouvé que les estimations des volumes des nodules sur des données compressées avec pertes (par SPIHT 3D) ne présentaient aucune variation significative, si l'on compare aux volumes connus dans la base de référence. Ces résultats bien que préliminaires sont encourageants et ouvrent des perspectives en compression avec pertes des scanners de type MDCT de poumons.

Enfin nous avons proposé des solutions pour réduire la complexité de notre schéma de compression à travers des outils analytiques dédiés à l'allocation de débits, étape incontournable des schémas dits "intra-bandes" comme le nôtre. La première allocation consacrée à la QVAZM 2D s'appuie sur un ajustement des courbes débit-distorsion en les approximant par un simple modèle exponentiel. Les paramètres de ce modèle sont calculés à partir d'un nombre réduit d'appels aux données. Ces courbes exponentielles conduisent à une résolution analytique de l'allocation de débits

négligeable en terme de complexité. Nous avons obtenu des performances proches (en termes de qualité) de celles obtenues par une approche lagrangienne mais avec une complexité calculatoire considérablement réduite. La seconde allocation proposée pour les images naturelles et médicales 3D a utilisé des modèles statistiques par blocs pour modéliser les courbes débit-distorsion. Les agglutinements de coefficients d'ondelettes significatifs se traduisent par une distribution de la norme des vecteurs dont la forme, comportant un mode très proche de zéro, ne peut être modélisée sous l'hypothèse i.i.d. Nous avons prouvé que la distribution de la norme des vecteurs de coefficients d'ondelettes se modélisait avec une grande précision par un mélange de lois gamma. Cette propriété a débouché sur la modélisation de la distribution conjointe des vecteurs de coefficients d'ondelettes par un mélange de gaussiennes généralisées (multidimensionnelles). Ce travail a permis de déduire des modèles de débit et de distorsion très précis qui débouchent sur une procédure d'allocation des débits avec un coût calculatoire réduit puisqu'il n'y a quasiment pas d'appel aux données. Les résultats numériques sont encore plus proches d'une QVAZM avec allocation lagrangienne. Encore une fois, les propriétés visuelles de la QVAZM (préservation des structures fines) sont conservées avec cette allocation. Par ailleurs, la flexibilité du modèle MMGG a permis de l'utiliser avec une grande précision sur les sous-bandes d'une pile d'images médicales (scanner ORL). Nous disposons ainsi d'une procédure d'allocation de débits compatible avec la contrainte de complexité médicale. Enfin, la précision du modèle MMGG et ses propriétés mathématiques avantageuses sont encourageantes pour envisager des applications dans d'autres disciplines du codage (vidéo, objet 3D...) ou pour d'autres problèmes du traitement du signal comme le débruitage.

Pour conclure, ce manuscrit a investigué un sujet très peu étudié à notre connaissance : **la compression avec pertes des images médicales**. Ce travail a montré que **sous certaines conditions**, la compression avec pertes des images volumiques radiologiques était **possible,** offrant ainsi des gains de compression significatifs par rapport aux méthodes sans pertes. Il ouvre ainsi de nombreux champs pour l'avenir de ce type de compression dans le domaine médical. En ce qui nous concerne, ces résultats prometteurs de la compression avec pertes (dont ceux de la QVAZM 3D) nous encouragent à poursuivre nos collaborations dans ce domaine. Nous allons tester notre algorithme ainsi que d'autres méthodes pour évaluer l'incidence de la compression avec pertes sur différents traitements classiques en imagerie médicale (recalage, segmentation, mesure de volumétrie en radiothérapie). Par ailleurs, nous avons vocation à améliorer notre algorithme en lui insérant des fonctionnalités (telle que le codage sans perte, l'amélioration de la qualité par raffinement, ou encore les régions d'intérêt).

Bibliographie

[1] *"ACR Technical standard for teleradiology"*, http ://www.unifesp.br/dis/set/disciplina/ materialdeapoio/ACRTechnicalStandardforTeleradiology.pdf, 2002.

[2] M. Antonini, M. Barlaud, P. Mathieu et I. Daubechies, *"Image coding using wavelet transform"*, IEEE Trans. Image Proc., 1(2) pp. 205-220, avril 1992.

[3] S. Armato, M. Altman , P.J. La Riviere, *"Automated detection of lung nodules in CT scans : effect of image reconstruction algorithm"*. Med Phys, n° 30, vol. 3, pp. 461-472, Mars 2003.

[4] K. Awai, K. Murao, A. Ozawa, et al, *"Pulmonary Nodules at chest CT : effect of computer-aided diagnosis on radiologists' detection performance"*, Radiology, vol. 230, pp. 347-352, 2004.

[5] M. Barlaud, *"Wavelets in image communication"*, Advances in image communication, Elsevier, 1994.

[6] M. Barlaud, C. Labit, *"Compression et codage des images et des vidéos"*, Traité IC2, série traitement du signal et de l'image, Hermès sciences publications, 2002.

[7] A. Baskurt, M. Khamadja, O. Baudin, F. Dupont, R. Prost, D. Revel, and R. Goutte, "*Adaptive image compression scheme. Application to medical images and diagnostic quality assessment*", Journal Européen des Systèmes Automatisés JESA, vol. 31, n° 7, pp. 1155-1171, Dec. 1997.

[8] H. Benoit-Cattin, A. Baskurt, and R. Prost, *"3D medical image coding using separable 3D wavelet decomposition and lattice vector quantization"*, Signal Processing, vol. 59, pp. 139-153, juin 1997.

[9] G. Battail, *"Théorie de l'information, application aux techniques de communication"*, Masson, 1997.

[10] A. Bilgin, G. Zweig, and M.V. Marcellin, "*Three-dimensional image compression with integer wavelet transform*", Applied Optics, vol. 39, N°. 11, pp. 1799-1814, 2000.

[11] J.F. Bonnans, J.C. Gilbert, C. Lemaréchal et C. Sagastizabal, *"Optimisation numérique"*, Springer Verlag, 1997.

[12] R2 Technology, *"Computer Aided Detection for Multi-Slice CT : ImageChecker"*, "http ://www.r2tech.com/chest_mdct/products/index.php".

[13] A.R. Calderbank, I. Daubechies, W. Sweldens, and B.L. Yeo, "*Lossless image compression using integer to integer wavelet transforms*", In Proc. of the International Conference on Image Processing (ICIP), pp. 596-599, 1997.

[14] R. C. Calderbank, I. Daubechies,W. Sweldens, and B.-L. Yeo, *"Wavelet transforms that map integers to integers"*,J. Appl. Comput. Harmon.Anal., vol. 5, pp. 332-369, 1998.

[15] P. Le Callet et D. Barba, "*Modele de perception couleur : application a l'evaluation de la qualite*", Traitement du signal,vol. 21,pp. 461-477, 2004.

[16] P. Le Callet, C. Viard-Gaudin et D. Barba, *"A convolutional neural network approach for objective video quality assessment",* IEEE Transactions on Neural Networks,Septembre, 2006.

[17] M. Cagnazzo, T. André, M. Antonini et M. Barlaud, *"A model-based motion compensated video coder with JPEG2000"*, ICIP, pp. 2255-2258, 2004.

[18] F. Chen, Z. Gao et J. Villasenor, *"Lattice Vector Quantization of Generalized Gaussian Sources"*, IEEE Transactions on Information Theory, vol.43, pp. 92-103,1997.

[19] S. Cho, D. Kim, and W. A. Pearlman, "*Lossless compression of volumetric medical images with improved 3-D SPIHT algorithm*," Journal of Digital Imaging , vol. 17, N°. 1, pp. 57-63, mars 2004.

[20] David A. Clunie, *"Lossless Compression of Grayscale Medical Images - Effectiveness of Traditional and State of the Art Approaches"*, SPIE Medical Imaging, San Diego, February 2000.

[21] W. Cochran, *"Some methods for strengthening the common chi-square test"*, Biometrics,vol.10, pp. 417-451, 1954.

[22] R. Coifman and V. Wickerhauser, *"Entropy-based algorithms for best basis selection"*, IEEE Trans. Inform. Theory, vol. 38, pp. 713-718, mars 1992.

[23] J.H Conway et N.J.A. Sloane, *"Sphere packing, lattices and groups"*, Springer Verlag, 1988.

[24] Cosman, P.C. Tseng, C. Gray, R.M. Olshen, R.A. Moses, L.E. Davidson, H.C. Bergin, C.J. Riskin, *"Tree-structured vector quantization of CT chest scans : image quality and diagnostic accuracy"*,

[25] P. Cosman, R. Gray, and R. Olshen, *"Evaluating quality of compressed medical images : SNR, subjective rating, and diagnostic accuracy"* ,Proc. EEE, vol. 82, pp. 919-932, juin 1994.

[26] P.C. Cosman, K.O. Perlmutter, S.M. Perlmutter, *"Tree-structured vector quantization with significance map for wavelet image coding"*, Proc. IEEE Data compression Conf. (DCC), J.A. Storer and M. Cohn, Eds. Snowbird Utah, IEEE computer society press, mars 1995

[27] J.-P. Coquerez et S. Philipp, *"Analyse d'images : filtrage et segmentation"*, Masson, 1995.

[28] P. Cosman, R. Gray, and R. Olshen, *"Quality evaluation for compressed medical images : Fundamentals"*, Handbook of Medical Imaging, Processing and Analysis, I.Bankman, Ed. New York : Academic, 2000.

[29] P. Cosman, R. Gray, and R. Olshen, *"Quality evaluation for compressed medical images : Diagnostic Accuracy"*, Handbook of Medical Imaging, Processing and Analysis, I.Bankman, Ed. New York : Academic, 2000.

[30] P. Cosman, R. Gray, and R. Olshen, *"Quality evaluation for compressed medical images : Statistical issues"*, Handbook of Medical Imaging, Processing and Analysis, I.Bankman, Ed. New York : Academic, 2000.

[31] T. Cover and J. Thomas, "*Elements of Information Theory*", Wiley, New York, 1991.

[32] I. Daubechies, "Ten Lectures on Wavelets", vol. 61 of Proc. CBMS-NSF Regional Conference Series in Applied Mathematics. Philadelphia, PA : SIAM, 1992.

[33] I. Daubechies and W. Sweldens, "*Factoring wavelet transform into lifting steps*", J. Fourier Anal. Appl., vol. 41, N°. 3, pp. 247-269, 1998. Fourier Anal. Appl., vol. 41, N°. 3, pp. 247-269, 1998.

[34] Denecker K, Van Overloop J, Lemahieu I, "*An experimental comparison of several lossless image coders for medical images*", Proc. 1997 IEEE Data Compression Conference.

[35] Devroy L, *"Non-Uniform Random Variate Generation"*, Springer-Verlag, 1986.

[36] Bradley J.Erickson, *"Irreversible Compression of Medical Images"*, Society for computer applications in radiology, 2000.

[37] T.R. Fischer, *"A pyramid vector quantizer"*, IEEE Transactions on Information Theory, vol. 32, pp. 568-583, juillet 1986.

[38] T.R. Fischer, *"Geometric source coding and vector quantization"*, IEEE Transactions on Information Theory, vol. 35, pp. 137-145, janvier 1989.

[39] R.Fletcher et C.M. Reeves, *"Function minimization by conjugate gradients"*, Computer J., vol 7, pp. 149-154, 1964.

[40] E.D. Frimout, J. Biemond et R.L. Lagendijk, *"Forward rate control for MPEG recording"*, in Proc. od SPIE, Visual communication and image processing '93, Cambridge, MA, novembre 1993.

[41] Z. Gao, F. Chen, B. Belzer et J. Villasenor, *"A comparison of the Z^n, E^8 and Leetch lattices for image subband quantization"*, in Proc. IEEE Data Compression Conference, J.A. Storer and M. Cohn, Eds., Snowbird (UTAH), pp. 312-321, mars 1995.

[42] C.germain C., V.Breton , P.Clarysse, Y.Gaudeau, T.Glatard, E.Jeannot, Y.Legre, C.Loomis, I.Magnin, J.Montagnat, J.M.Moureaux , A.Osorio , X.Pennec et R.Texier, *" Grid-enabling medical image analysis"*, Journal of Clinical Monitoring and Computing, vol. 19, pp. 339-349, Octobre 2005.

[43] A. Gersho et R.M. Gray, "*Vector Quantization and Signal Compression*", Kluwer Academic Publishers, 1992.

[44] Goldberg MA, Gazelle GS, Boland GW, et al, *"Focal hepatic lesions : effect of three-dimensional wavelet compression on detection at CT"*. Radiology 1994, vol. 190, pp. 517-524.

[45] R. M. Gray et D. L. Neuhoff, *"Quantization"*, in IEEE Trans. on Information Theory, vol. 44, N°.6, octobre 1998.

[46] L. Guillemot, *"Une approche vectorielle pour exploiter le contenu de l'image en compression et tatouage",* Thèse de l'Université Henri Poincaré - Nancy, décembre 2004.

[47] L.Guillemot, Y.Gaudeau, J.-M. Moureaux *" A new fast bit allocation procedure for image coding based on discrete wavelets transform and dead zone lattice quantization",* ICIP 2005, Gêne, Italie, du 11 au 14 septembre, 2005.

[48] L.Guillemot, Y.Gaudeau, S. Moussaoui et J.-M. Moureaux, *" An Analytical gamma mixture based rate-distortion model for lattice vector quantization"*, EUSIPCO, Florence, septembre 2006.

[49] L.Guillemot, Y.Gaudeau, S. Moussaoui et J.-M. Moureaux, *"Lattice Vector Quantization Rate-Distortion Models and Codebooks dedicated to block mixture densities."*, Soumis à IEEE Signal Processing.

[50] D.A. Huffman, *"A method for the construction of minimum-redundancy codes"*, Proc. of the IRE, vol.40, pp. 1098-1101, septembre 1952.

[51] Huang and Schultheiss,*"Block Quantization of Correlated Gaussian Random Variables"*, IEEE Transactions on Communications Systems, pp. 289-296, 1963.

[52] R. Hunter et A.H. Robinson, *"Runlength"*, Proceedings of the IEEE, vol.68, pp. 854-867, 1980.

[53] D.G. Jeong et J.D. Gibson, *"Uniform and piecewise uniform lattice vector quantization for memoryless Gaussian and Laplacian sources"*, IEEE Transactions on Information Theory, vol. 39, pp. 786-804, mai 1993.

[54] A, Kalyanpur, V. Neklesa, C. Taylor, et al, *"Evaluation of JPEG and wavelet compression of body CT images for direct teleradiologic transmission"*, Radiology, vol. 217, pp. 772-779, 2000.

[55] J. Ko , and D. Naidich , *"Computer-aided diagnosis and the evaluation of lung disease"*, J Thoracic Imaging, vol. 19, pp. 136-155, 2004.

[56] J. Ko, H. Rusinek, D. Naidich, et al, *"Wavelet compression of low-dose chest CT data : effect on lung nodule detection"*, Radiology, vol. 228, pp. 70-75, 2004.

[57] B-J. Kim and W.A. Pearlman, "*An Embedded Wavelet Video Coder Using Three-Dimensional Set Partitioning in Hierarchical Trees*," IEEE Data Compression Conference, pp. 251-260, mars1997.

[58] Y. S. Kim and W. A. Pearlman, "*Lossless volumetric Image Compression,*" in Applications of Digital Image Processing XXII, Proceedings of SPIE vol. 3808, pp. 305-312, 1999 .

[59] Kivijärvi J, et al, "*A comparison of lossless compression methods for medical images*", Computerized Medical Imaging and Graphics, 22, pp 323-339, 1998.

[60] F. Li, Sone S. Takashima, et al, *"Effects of JPEG and wavelet compression of spiral low-dose CT images on detection of small lung cancers"*. Acta Radiol, vol. 42, pp. 156-160, 2001.

[61] Y. Linde, A. Buzo et R.M. Gray, *" An algorithm for Vector Quantizer design"*, IEEE Transactions on communications, vol. COM-28, N°.1, pp. 84-95, janvier 1980.

[62] J. Liu et P. Moulin, *"Complexity-Regularized Image Denoising"*, IEEE Trans Image Processing, N°. 10, pp. 841-851, 2001.

[63] S. LoPresto, K. Ramchadran and M.T.Orchard, "Image coding based on mixture modeling of wavelet coefficients and a fast estimation-quantization framework", in Data Compression Conference '97, Snowbird Utah, pp. 221-230, 1997

[64] P. Loyer, J.-M. Moureaux et M. Antonini, *"Lattice codebook enumeration for generalized gaussian source"*, in IEEE Transactions on Information Theory, vol. 49, N°. 2, pp. 521-528, février 2003.

[65] S. Mallat, *"A theory for multiresolution signal decomposition : the wavelet representation"*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 11, pp. 674-693, juillet 1989.

[66] S. Mallat, *"Une exploration des signaux en ondelettes"*, les éditions de l'école polytechniques, 2000.

[67] M.W. Marcellin, M.J. Gormish, A. Bilgin, and M. P. Boliek, *"An overview of JPEG-2000"*, in Proc. DCC 2000, Snowbird, UT, pp. 523-541, mars 2000.

[68] J.R. Mayo, K. Kim, S. MacDonald , et al. *"Reduced radiation dose helical chest CT : Effect on reader evaluation of structures and lung findings"*, vol. 232, pp. 749-756, Radiology 2004.

[69] Gloria. Menegaz and J.-Ph. Thiran, *Object-based Coding of 3D Medical Data Without Boundary Artifacts*, Transactions on Image Processing, 2002.

[70] Gloria Menegaz et Jean-Philippe Thiran, *"Lossy to Lossless Object-Based Coding of 3-D MRI Data"*, IEEE Transactions on Image Processing vol.11, septembre 2002.

[71] S.-G. Miaou et S.-T. Chen, *"Automatic quality control for wavelet-based compression of volumetric medical images using distortion-constrained adaptive vector quantization"*, IEEE Transactions on Medical Imaging, vol.23, novembre 2004.

[72] Gloria Menegaz et Jean-Philippe Thiran, *"3D Encoding/2D Decoding of Medical Data "*, IEEE Transactions on Medical Imaging, vol.22, mars 2003.

[73] A. Mostefaoui, F. Prêteux, V. Lecuire et J. M. Moureaux, *"Gestion des données multimédias"*, traité IC2 Information - Commande - Communication, hermès sciences publications, 2004.

[74] J.M. Moureaux, M. Antonini, et M. Barlaud, "*Counting lattice points on ellipsoids : Application to image coding*", Electronic Letters., vol. 31, pp. 1224-1225, juillet 1995.

[75] J.M. Moureaux, P. Loyer et M. Antonini, *"Low-Complexity Indexing Method for Z^n and D^n lattice quantizers"*, IEEE Transactions on Communications, vol. 46, 12, pp. 1602-1609, décembre 1998.

[76] S. Moussaoui, D. Brie, A. M. Mohammad-Djafari et C. Carteret *"Separation of Non-negative Mixture of Non-negative Sources using a BayesianApproach and MCMC Sampling"*, IEEE Transactions on Signal Processing, vol. 54, pp. 4133 - 4155, novembre 2006.

[77] R. Ochs, E. Angel, K. Boedeker et al, *"The influence of CT dose and reconstruction parameters on automated detection of small pulmonary nodules"*, Medical Imaging, Proceedings of SPIE, Vol. 6144, 2006.

[78] A. Ortega et K. Ramchadran, "Rate-distorsion methods for image and video compression", IEEE Signal Processing Magazine, pp. 23-50, novembre, 1998.

[79] C. Parisot et M. Antonini, "*Compression d'images satellites haute résolution par quantification vectorielle algébrique et allocation de débits optimale*", I3S Laboratory internal report n°9819, Université de Nice Sophia Antipolis (France), octobre 1998.

[80] C. Parisot, *"Allocations basées modèles et transformées en ondelettes au fil de l'eau pour le codage d'images et de vidéos"*, Thèse de l'université de Nice - Sophia Antipolis, 2003.

[81] D. Peel et G. MacLahlan, *"Finite Mixture Models"*, Wiley interscience, 2000.

[82] M Penedo et al, *" Free-response receiver operating characteristic evaluation of lossy JPEG2000 and object-based set partitioning in hierarchical trees compression of digitized mammograms "*, Radiology, vol. 237, N°2, pp. 450-457, 2005.

[83] W.Philips, S. Van Assche, D De Rycke et K Denecker, *"State of-the-art for lossless compression of 3D medical images sets"*, Computerized Medical Imaging and Graphics, 2001.

[84] W. H. Press, S. A. Teukolsky, W. T. Vetterling et B. P. Flannery, *"Numerical Recipies in C"*, Cambridge university press.

[85] A. Przelaskowski, "*Vector quality measure of lossy compressed medical imagesVector quality measure of lossy compressed medical images*", Computers in Biology and Medecine, 2003

[86] P. Raffy, *"Modélisation, optimisation et mise en œuvre de quantificateurs bas débits pour la compression d'images utilisant une transformée en ondelettes"*, Thèse de l'Université de Nice - Sophia Antipolis, décembre 1997.

[87] P. Raffy, M. Antonini et M. Barlaud, *"A new optimal subband bit allocation procedure for very low bit rate image coding"*, IEE Electronics Letters, vol. 34, 7, p 647, avril 1998.

[88] P. Raffy, A. Najmi et R. Gray, *"Blind denoising using a wavelet coder"*. Proc IEEE 33rd Asilomar Conference on Signals Systems & Computers, pp. 1282-1286, 1999.

[89] P. Raffy, M. Antonini et M. Barlaud, *"Distortion-Rate Models for Entropy-Coded Lattice Vector Quantization"*, IEEE Transactions on Image Processing, vol. 9, pp. 2006-2017, décembre 2000.

[90] P. Raffy, Y. Gaudeau, D. P. Miller, et J-M. Moureaux, "*Computer Aided Detection (CAD) of Solid Lung Nodules in Lossy Compressed MDCT Chest Exams*", ECR, Vienne, mars 2006.

[91] P. Raffy, Y. Gaudeau, D. P. Miller, et J-M. Moureaux, "*Effect of 3D Wavelet Image Compression on Computer Aided Detection (CAD) Lung Nodule volumetry*", ECR, Vienne, mars 2006.

[92] P. Raffy, Y. Gaudeau, D. P. Miller, et J-M. Moureaux, *"Computer Aided Detection of Solid Lung Nodules in Lossy Compressed Computed Tomography Chest Exams"*, Academic Radiology. Accepté.

[93] P. Rault et C. Guillemot, *"Lattice vector quantization with reduced or without look-up table"*, Proc. SPIE Electronic Imaging, Santa Clara, Fl, janvier 1998.

[94] J. Reichel, G. Menegaz, M. Nadenau et M. Kunt, "*Integer Wavelet Transform for Embedded Lossy to Lossless Image Compression",* IEEE Transactions on Image Processing, vol. 10, N°. 3, pp. 383-392, mars 2001.

[95] J. Rissanen, "*Universal coding, information, prediction, and estimation*", IEEE Trans. Inf. Theory 30, 629-636, 1984.

[96] J. Roca, M. Antonini et M. Barlaud, *"Optimisation de quantificateurs vectoriels à entropie contrainte pour le codage d'images multirésolution"*, $15^{ème}$ colloque GRETSI, pp. 839-842, Juan les Pins, 1995.

[97] H. Rusinek, D. Naidich, G. McGuinness, et al, *"Pulmonary nodule detection : low-dose versus conventional CT"*, Radiology, vol.209, pp. 243-249, 1998.

[98] A. Said et W. Pearlman,"*A new, fast, and efficient image codec based on set partitioning in hierarchical trees*",IEEE Transactions Circuits Syst. Video Technol., vol. 6, pp. 243-250, juin 1996.

[99] A. Said et W. Pearlman,"*An Image Multiresolution Representaion for Lossless and Lossy Compression*",IEEE Trans. on Image Processing, vol. 5, pp. 1303-1310, septembre. 1996.

[100] J. Sajous, A. Megibow, H. Rusinek, C. Ladneret al, *"A wavelet-based 3D lossy JPEG 2000 algorithm allows 50-fold compression of CT image data preserving detection and measured volume of small hepatic metastases"*. Radiology, vol. 539, 2004.

[101] P.Schelkens, J Barbarien et J.Cornelis, *"Compression of volumetric medical data based on cube-splitting"*, Proc SPIE Compression on Applications of Digital Image Processing XXIII, vol 4115, pp. 91-101, août 2000.

[102] P. and A. Munteanu and J. Cornelis, "*Wavelet Coding of volumetric Medical Data",*Proc. 3rd IEEE Benelux Signal Processing Symposium, Louvain, Belgique, mars 2002.

[103] P. Schelkens, A. Munteanu, J. Barbarien, M. Galca,X.Giro-Nieto et J. Cornelis,"*Wavelet Coding of volumetric Medical Datasets",* IEEE Transactions on Medical Imaging, vol 22, mars 2003.

[104] J. Shapiro, "*Embedded image coding using zerotrees of wavelet coefficients*," IEEE Transactions Signal Processing, vol. 41, pp. 3445-3462, Dec. 1993.

[105] W. Sweldens, "*The lifting scheme : A construction of second generation wavelet*", SIAM J. Math. Anal., vol. 29, pp. 511-546, 1997.

[106] D. Taubman, "*High performance scalable image compression with EBCOT*", IEEE Transactions on Image Processing, vol. 9, N°. 7, pp. 1158-70, juin 2000.

[107] D. Taubman, W.M. Marcellin *"JPEG2000 image compression fundamentals, standards and practice"*, Kluwer Academic Publishers, 2002.

[108] David Taubman, Erik Ordentlich, Marcelo Weinberger et Gadiel Seroussi, "*Embedded block coding in JPEG 2000*", Signal Processing : Image Communication 17 (1) pp. 49-72, 2002.

[109] M.J. Tsai J. Villasenor et F. Chen, *"Stack-Run image coding"*, IEEE Transactions Circuits Syst. Video Technol., vol. 6, pp. 519-521, octobre 1996.

[110] T. Voinson, L Guillemot et J.M. Moureaux, "*Image compression using Lattice Vector Quantization with code book shape adapted thresholding*", IEEE 2002 International Conference on Image Processing, Rochester, New York, septembre 2002.

[111] T. Voinson, *" Quantification Vectorielle Algébrique avec Zone Morte : application à la compression d'images à bas débit et au tatouage d'images"*, Thèse de l'Université Henri Poincaré - Nancy, février 2003.

[112] J. Wang and H. K. Huang, *"Medical image compression by using three-dimensional wavelet transformation"*, IEEE Transactions on Medical Imaging, vol.15, 547-554, août 1996.

[113] S. Wong, L. Zaremba, D. Gooden, and H. Huang, *"Radiologic image compression-a review"*, Proc. IEEE, vol. 83, pp. 194-219, février 1995.

[114] M. Weinberger, G. Seroussi, G.Sapiro, "*The LOCO-I lossless image compression algorithm : Principles and standardization into JPEG-LS*", Technical Report HPL-98-193, HP Computer Systems Laboratory, novembre 1998. http ://www.hpl.hp.com/techreports/98.

[115] X. Wu and J.-H. Chen, "*Context modeling and entropy coding of wavelet coefficients for image compression*", in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal, pp. 3097-3100, New York, 1997.

[116] X.Wu , "*Lossless compression of continuous-tone images via context selection, quantization and modeling*", IEEE Transactions on Image Processing, vol. 6, pp. 656-664, 1997.

[117] J. Xu, Z. Xiong, S. Li, and Y.-Q. Zhang, "*3-D embedded subband coding with optimal truncation (3-D ESCOT)*", J. Appl .Computat. Harmonic Anal., vol. 10, pp. 290-315, mai 2001.

[118] Z. Xiong, X. Wu, S. Cheng et J. Hua, "*Lossy-to-Lossless Compression of Medical volumetric Data Using Three-dimensional Integer Wavelet Transforms"*, IEEE Transactions on Medical Imaging, vol.22, mars 2003.

[119] Zalis ME, Hahn PF, Arellano RS, et al, *"CT colonography with teleradiology : effect of lossy wavelet compression on polyp detection - initial observations"*, Radiology, vol. 220, pp. 387-392, 2001.