Conception et évaluation d'un nouvel algorithme de reconstruction itérative en tomodensitométrie à faisceau conique implanté sur matériel graphique

Thèse

Dmitri Matenine

Doctorat en physique
Philosophiæ doctor (Ph. D.)

Québec, Canada

Conception et évaluation d'un nouvel algorithme de reconstruction itérative en tomodensitométrie à faisceau conique implanté sur matériel graphique

Thèse

Dmitri Matenine

Sous la direction de:

Philippe Després, directeur de recherche
Yves Goussard, codirecteur de recherche

Résumé

La présente thèse s'inscrit dans le domaine de la physique médicale et, plus précisément, de l'imagerie médicale tridimensionnelle (3D) et de la dosimétrie 3D pour la radiothérapie. L'objectif global du travail était de concevoir et évaluer un nouvel algorithme de reconstruction itératif rapide pour la tomodensitométrie (TDM) à faisceau conique, une modalité consistant à créer des images 3D des densités du sujet imagé à partir de mesures d'atténuation partielle d'un faisceau de radiation incidente. Cet algorithme a été implanté sur matériel graphique (GPU), une plate-forme de calcul hautement parallèle, menant à la conception de stratégies d'optimisation originales. En premier lieu, un nouvel algorithme itératif statistique régularisé, dénommé OSC-TV, a été conçu et implanté sur GPU. Il a été évalué sur des ensembles de projections synthétiques et cliniques de TDM à rayons X à faisceau conique. L'algorithme proposé a démontré une qualité d'image supérieure à celle de méthodes semblables pour des acquisitions basse-dose, ainsi que des temps de reconstruction compatibles avec les activités cliniques. L'impact principal de ce travail est la capacité d'offrir au patient une réduction de dose de radiation ionisante de deux à quatre fois par rapport aux protocoles d'acquisition usuels. En second lieu, cet algorithme a été testé sur des données expérimentales en tomographie optique à faisceau conique, donnant lieu à l'une des premières études de ce genre. La résolution spatiale des images 3D résultantes a été améliorée et le bruit a été réduit. L'on a aussi démontré l'importance de considérer le spectre de la source lumineuse afin d'assurer la justesse de l'estimation des densités. Le principal impact de l'étude est la démonstration de la supériorité de la reconstruction itérative pour des données affectées par les aberrations propres à la tomographie optique à faisceau conique, résultant potentiellement en l'amélioration de la dosimétrie 3D par gel radiochromique en radiothérapie. En troisième lieu, différentes approches de gestion de la matrice-système de type exact à rayons fins ont été évaluées pour la TDM à faisceau conique. Le pré-calcul et le stockage complet de la matrice-système dans la mémoire vive du GPU s'est montré comme l'approche la plus rapide, mais la moins flexible en termes de géométries représentables, en raison de la taille limitée de la mémoire vive. Le traçage de rayons à la volée est apparu très flexible, offrant aussi des temps de reconstruction raisonnables. En somme, les trois études ont permis de mettre en place et d'évaluer la méthode de reconstruction proposée pour deux modalités de tomographie, ainsi que de comparer différentes façons de gérer la matrice-système.

Abstract

This thesis relates to the field of medical physics, in particular, three-dimensional (3D) imaging and 3D dosimetry for radiotherapy. The global purpose of the work was to design and evaluate a new fast iterative reconstruction algorithm for cone beam computed tomography (CT), an imaging technique used to create 3D maps of subject densities based on measurements of partial attenuation of a radiation beam. This algorithm was implemented for graphics processing units (GPU), a highly parallel computing platform, resulting in original optimization strategies. First, a new iterative regularized statistical method, dubbed OSC-TV, was designed and implemented for the GPU. It was evaluated on synthetic and clinical X ray cone beam CT data. The proposed algorithm yielded improved image quality in comparison with similar methods for low-dose acquisitions, as well as reconstruction times compatible with the clinical workflow. The main impact of this work is the capacity to reduce ionizing radiation dose to the patient by a factor of two to four, when compared to standard imaging protocols. Second, this algorithm was evaluated on experimental data from a cone beam optical tomography device, yielding one of the first studies of this kind. The spatial resolution of the resulting 3D images was improved, while the noise was reduced. The spectral properties of the light source were shown to be a key factor to take into consideration to ensure accurate density quantification. The main impact of the study was the demonstration of the superiority of iterative reconstruction for data affected by aberrations proper to cone beam optical tomography, resulting in a potential to improve 3D radiochromic gel dosimetry in radiotherapy. Third, different methods to handle an exact thin-ray system matrix were evaluated for the cone beam CT geometry. Using a GPU implementation, a fully pre-computed and stored system matrix yielded the fastest reconstructions, while being less flexible in terms of possible CT geometries, due to limited GPU memory capacity. On-the-fly ray-tracing was shown to be most flexible, while still yielding reasonable reconstruction times. Overall, the three studies resulted in the design and evaluation of the proposed reconstruction method for two tomographic modalities, as well as a comparison of the system matrix handling methods.

Table des matières

Liste des tableaux

Liste des figures

Liste des abréviations

1D	Mono-dimensionnel(le)
2D	Bi-dimensionnel(le)
3D	Tri-dimensionnel(le)
4D	Quadri-dimensionnel(le)
ADC	convertisseur analogique-digital (analog to digital converter)
ALU	Unité de calcul arithmétique et logique (arithmetic logic unit)
ART	Technique de reconstruction algébrique (algebraic reconstruction technique)
BFGS	Méthode d'optimisation Broyden–Fletcher–Goldfarb–Shanno
CBCT	Tomodensitométrie à faisceau conique (cone beam computed tomography)
CIP	Calcul informatique de pointe
CHU	Centre hospitalier universitaire
CMOS	Technologie de fabrication *complementary metal–oxide–semiconductor*
CNR	Rapport contraste sur bruit (contrast to noise ratio)
CPU	Processeur central (central processing unit)
CRSNG	Conseil de recherches en sciences naturelles et en génie du Canada
CT	Tomodensitométrie (computed tomography)
DD	Méthode guidée par la distance (distance-driven method)
DEL	Diode électroluminescente
EM	Espérance-maximisation (expectation-maximization)
ERF	Fonction de réponse à une marche (edge response function)
FDK	Algorithme de rétroprojection filtrée de Feldkamp, Davis et Kress
FDP	Fonction de densité de probabilité
FOV	Champ de vision (field of view)
FONCER	Programme de formation orientée vers la nouveauté, la collaboration et l'expérience en recherche
FRQ	Fonds de recherche du Québec
FWHM	Pleine largeur à mi-hauteur (full width at half-maximum)
GFLOP	Milliards (giga-) d'opérations en virgule flottante

GPGPU	Calcul d'usage général sur matériel graphique (general purpose GPU [computing])
GPU	Matériel graphique (graphics processing unit)
HU	Unités Hounsfield (Hounsfield units)
ICD	Méthode de la descente de coordonnées itérative (iterative coordinate descent)
ID	Identifiant unique d'un processus léger
IR	Iterative reconstruction
IMRT	Intensity-modulated radiotherapy
IRM	Imagerie par résonance magnétique
LED	Diode électroluminescente (light-emitting diode)
MAP	Maximum *a posteriori*
MCar	Moindres carrés
MLEM	Maximisation de la vraisemblance par espérance-maximisation (maximum likelihood expectation-maximisation)
MPRTN	Medical Physics Research Training Network
MTF	Fonction de transfert de modulation (modulation transfer function)
MV	Maximisation de la vraisemblance
NLM	Régularisation par sous-images (non-local means regularization)
NRMSD	Différence quadratique normalisée (normalized root-mean-square difference)
NURBS	B-splines rationnelles non uniformes (non-uniform rational basis splines)
OD	Densité optique (optical density)
OSC	Méthode convexe avec sous-ensembles ordonnés (ordered subsets convex method)
OTF	Calcul à la volée (on-the-fly)
PCI	Bus local ou interne d'un ordinateur (Peripheral Component Interconnect)
PL	Processus léger
POCS	Projections sur les ensembles convexes (projections onto convex sets)
PMMA	Polyméthacrylate de méthyle
PMT	Tube photomultiplicateur (photomultiplier tube)
RAM	Mémoire vive (random access memory)
RCMI	Radiothérapie conformationnelle avec modulation d'intensité
RI	Reconstruction itérative
ROI	Région d'intérêt (region of interest)
SART	Technique de reconstruction ART avec mises à jour simultanées
SIRT	Technique de reconstruction itérative avec mises à jour simultanées
SNR	rapport signal sur bruit (signal to noise ratio)

SF	Méthode d'empreintes séparables (separable footprints method)
SM	Multiprocesseur de calcul sur GPU (streaming multiprocessor)
SQS	Fonctions de substitution quadratiques séparables (SQS) (separable quadratic surrogates)
SIMD	Instruction unique, données multiples (single-instruction, multiple data)
TDM	Tomodensitométrie
TF	Transformée de Fourier
TV	Minimisation de la variation totale (total variation minimization)

À ma chère épouse Elena

Remerciements

Les travaux présentés dans cette thèse sont le fruit d'une étroite collaboration et ont été faits grâce au support de nombreuses personnes.

D'abord, je tiens à remercier mon directeur de recherche Philippe Després, qui a guidé la démarche scientifique de ce projet. J'aimerais en particulier souligner sa patience et compréhension qui ont su m'encourager, ainsi que son intuition académique qui m'a aidé à prendre des parcours fructueux.

Je tiens également à remercier Yves Goussard, le codirecteur de cette recherche. Sa rigueur scientifique et son esprit analytique m'ont amené à former un regard critique sur ce travail et d'améliorer la cohésion de la démarche scientifique, ainsi que la clarté de la présentation des résultats.

Un grand merci à Julia Mascolo-Fortin qui s'est impliquée dans le développement et les tests du logiciel de reconstruction utilisé lors de ces travaux, ainsi que dans les expériences de tomographie optique. Il en va de même pour Geoffroi Côté, qui a contribué au logiciel.

J'aimerais également remercier les professeurs et collègues au groupe de recherche en physique médicale : Louis Archambault, Luc Beaulieu, Jonathan Boivin, Songye Cui, Marie-Ève Delage, Patricia Duguay-Drouin, Romain Espagnet, Cédric Laliberté-Houdeville, Jean-François Montégiani, et plusieurs autres, dont les conseils pratiques et le support moral ont été fort appréciés.

Je tiens à reconnaître le support financier et technique de différentes institutions : le Fonds de recherche du Québec—Nature et technologies (FRQ-NT), le Medical Physics Research Training Network (MPRTN), lui-même soutenu par le Programme de formation orientée vers la nouveauté, la collaboration et l'expérience en recherche (FONCER) du Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG), ainsi que l'Université Laval, le CHU de Québec et l'École Polytechique de Montréal.

J'aimerais remercier John Miller et Jen Dietrich de Modus Medical Devices Inc. pour leurs précieux conseils lors des travaux réalisés sur le système de tomographie optique DeskCAT™.

Enfin, je remercie mon épouse, Elena, mes parents Pavel et Irina, ma soeur Ekaterina et mes

beaux-parents, Anatole et Galina, d'avoir été à mes côtés durant ce long parcours et d'avoir bien pris soin de moi. Un merci spécial pour ma fille Ksenia et mon fils Luka.

Avant-propos

Cette thèse inclut trois articles qui ont été publiés ou soumis à des journaux scientifiques. L'information sur chacun des auteurs ainsi que leurs contributions respectives est indiquée ci-après. Dmitri Matenine est le premier auteur de chacun de articles.

Chapitre 2 : GPU-Accelerated Regularized Iterative Reconstruction for Few-view Cone Beam CT

Dmitri Matenine[1], Yves Goussard[2] et Philippe Després[1,3,4]

Publié : Medical Physics **42** (4) 1505-1517, Avril 2015

Contributions : Dmitri Matenine a effectué la conception de l'algorithme de reconstruction, a effectué la programmation CPU/GPU de l'algorithme de reconstruction, a généré et/ou mis en forme les données d'entrée, a généré les résultats bruts et les résultats de synthèse, a rédigé l'article et a corrigé l'article en fonction des commentaires des co-auteurs, de l'éditeur du journal et des arbitres désignés. Yves Goussard a participé à la classification de l'algorithme proposé par rapport aux approches disponibles dans la littérature et à la révision de l'article. Philippe Després a participé à l'élaboration de la démarche expérimentale, la programmation du module de traçage de rayons, la mise en forme des données d'entrée, l'interprétation des résultats et la révision de l'article.

Chapitre 3 : Evaluation of the OSC-TV iterative reconstruction algorithm for cone-beam optical CT

Dmitri Matenine[1], Julia Mascolo-Fortin[1], Yves Goussard[2] et Philippe Després[1,3,4]

Publié : Medical Physics **42** (11) 6376-6386, Novembre 2015

Contributions : Dmitri Matenine a effectué l'acquisition des données expérimentales, a mis en forme les données d'entrée, a généré les résultats bruts et les résultats de synthèse, a rédigé l'article et a corrigé l'article en fonction des commentaires des co-auteurs, de l'éditeur du journal et des arbitres désignés. Julia Mascolo-Fortin a participé à l'acquisition des données expérimentales et la révision de l'article. Yves Goussard a participé à l'interprétation des résultats et la révision de l'article. Philippe Després a participé à l'élaboration de la démarche

expérimentale, l'interprétation des résultats et la révision de l'article.

Chapitre 4 : System matrix computation vs storage on GPU : a comparative study in cone beam CT

Dmitri Matenine[1], Geoffroi Côté[1], Julia Mascolo-Fortin[1], Yves Goussard[2] et Philippe Després[1,3,4]

Soumis au journal Medical Physics le 16 mars 2017

Contributions : Dmitri Matenine a effectué la conception de l'architecture du programme, la programmation des modules de gestion de la mémoire, a généré et mis en forme les données d'entrée, a généré les résultats bruts et les résultats de synthèse, a rédigé l'article et a corrigé l'article en fonction des commentaires des co-auteurs. Geoffroi Côté a participé à la programmation du module de traçage de rayons. Julia Mascolo-Fortin a participé à la conception des modules d'entrées/sorties du programme, aux tests de robustesse du programme et à la révision de l'article. Yves Goussard a participé à l'interprétation des résultats et la révision de l'article. Philippe Després a participé à la programmation du module de traçage de rayons, l'interprétation des résultats et la révision de l'article.

Affiliations des auteurs :

1. Département de physique, de génie physique et d'optique, Université Laval, Québec, Québec, G1V 0A6, Canada
2. Département de génie électrique / Institut de génie biomédical, École Polytechnique de Montréal, C.P. 6079, succ. Centre-ville, Montréal, Québec H3C 3A7, Canada
3. Centre de recherche sur le cancer, Université Laval, Québec, Québec G1V 0A6, Canada
4. Département de radio-oncologie et Centre de recherche du CHU de Québec, Québec, Québec G1R 2J6, Canada.

Chapitre 1

Introduction

La présente recherche se situe dans le domaine de la physique médicale, une discipline qui utilise différents phénomènes physiques pour le bénéfice des patients, via l'élaboration et le perfectionnement de diverses techniques de diagnostic et de thérapie. Plus précisément, les travaux décrits ici touchent à l'imagerie médicale. Cette recherche fait également appel à des notions de traitement d'images numériques, ainsi qu'au calcul informatique de pointe (CIP). Les résultats de cette recherche s'appliquent à la modalité d'imagerie appelée *tomodensitométrie* (TDM). Le terme provient du mot grec *tomos*, qui signifie *tranche* et *densitométrie*, c.-à-d. mesure de densité. Il s'agit d'une technique d'imagerie volumétrique (3D) qui permet de visualiser l'intérieur du sujet imagé *tranche par tranche*. Les estimations de densité sont basées sur les mesures de l'atténuation d'un rayonnement pénétrant auquel le sujet est partiellement transparent. La TDM est dite *non-invasive* pour le patient et *non-destructive* lorsqu'un objet inanimé est imagé. La TDM est effectuée sur des patients en utilisant des rayons X et constitue actuellement un service incontournable offert par les départements de radiologie diagnostique des hôpitaux publics, ainsi que dans des cliniques privées. Cette modalité est reconnue pour des acquisitions rapides, de l'ordre de 1 min et une résolution spatiale de l'ordre de 1 mm, ce qui en fait un outil indispensable pour évaluer une large éventail de problèmes de santé. La recherche scientifique contribue à l'amélioration de divers aspects de la TDM, dont la qualité d'image, la réduction de la dose de rayons X associée à cet examen, ainsi que le potentiel d'imagerie quantitative possible via la TDM. La présente thèse s'inscrit dans cette optique, et les sections suivantes articulent les problématiques ayant motivé la recherche et l'impact potentiel de cette dernière.

1.1 Tomodensitométrie à rayons X

La tomodensitométrie (TDM) ou CT – de l'anglais *computed tomography* – est une extension de la radiographie classique vers l'imagerie 3D. La TDM nécessite une transformation mathématique afin d'obtenir une image 3D du patient à partir de mesures d'atténuation de la

radiation qui le traverse, par exemple, à partir de plusieurs images radiographiques planaires. Sa réalisation pratique a été possible grâce à l'avènement des premiers ordinateurs. Le premier prototype fonctionnel d'un tomodensitomètre a été proposé en 1972 par Sir Godfrey Hounsfield, un ingénieur britannique. Sir Hounsfield a obtenu le Prix Nobel de la médecine de 1979 pour son invention, partagé avec Allan McLeod Cormack qui a effectué des travaux similaires indépendamment.

La tomodensitométrie requiert des calculs pour obtenir l'image du patient, c'est-à-dire qu'une reconstruction de l'image à partir de données brutes est nécessaire. En fait, il s'agit de solutionner un problème inverse, qui se base sur un modèle mathématique de l'atténuation des rayons X dans le corps du patient. En général, plus le modèle du problème est proche de la réalité physique, plus la solution est juste et plus la complexité calculatoire est élevée. La performance des ordinateurs, c'est-à-dire la vitesse de calcul et la capacité de la mémoire étant finies, il est crucial de choisir un modèle qui propose un bon équilibre entre le temps de calcul et la justesse de l'image.

Une autre préoccupation importante en TDM est liée aux dangers de la radiation ionisante. La quantification du risque de cancer radio-induit lié aux examens de TDM est à ce jour un sujet de recherche actif et aussi une source de controverse chez les experts en radiologie [1,2]. Toutefois, il existe un consensus important et durable : l'imagerie TDM devrait être effectuée suite à une indication médicale valide et la dose devrait être aussi basse que raisonnablement possible pour assurer un diagnostic concluant [1,3]. Cependant, une grande partie des appareils actuels utilisent des algorithmes de reconstruction relativement simples, comme la rétroprojection filtrée de Feldkamp Davis et Kress (FDK) [4]. Cet algorithme requiert des données peu bruitées acquises à des doses relativement élevées pour offrir des images de qualité. Il existe une classe d'algorithmes avancés, dits *itératifs*, qui améliorent graduellement l'estimation de l'image 3D au prix de volumes de calculs nettement plus élevés. Ces derniers sont plus robustes face à des projections bruitées acquises avec moins de photons X [5–8].

La reconstruction tomographique d'un objet tridimensionnel est possible à partir d'un ensemble de projections planaires (2D) de cet objet, aussi appelée *sinogramme*. Plus précisément, la tomodensitométrie est basée sur des images planaires obtenues par l'atténuation partielle des rayons X traversant le patient. L'énergie des photons en radiologie se situe dans l'intervalle approximatif de 30 keV à 140 keV. Ce choix est expliqué en détail à la section suivante. Les projections planaires sont obtenues en plaçant le sujet entre une source de rayons X et un détecteur. La source en milieu médical est généralement un tube à rayons X, alors que des sources radioactives sont souvent employées en imagerie non-destructive industrielle. Différents types de détecteurs ont été utilisés historiquement, avec la motivation de réduire le temps d'acquisition et d'augmenter la résolution spatiale des images. Le premier appareil fabriqué par Hounsfield en 1972 a été muni d'un tube photomultiplicateur (PMT) unique [9]. Son prototype a donnée naissance au premier appareil commercial produit en série : le scanner

d'EMI (Grande Bretagne). L'acquisition d'une seule projection nécessitait alors une translation orthogonale à l'axe source-détecteur et ces appareils ont été désignés comme étant de première génération. Vers 1974 sont apparus des appareils munis d'une *barrette* de 30 PMT pour réduire le nombre de translations par projection [10]. Ceci a marqué la deuxième génération de la TDM médicale. Durant les années 1980 est apparue la géométrie du *faisceau en éventail*, qui permettait d'acquérir une projection complète sans translation, car le détecteur couvrait le champ de vue complet. À ce moment, la barrette de détecteurs était constituée de chambres à ionisation au xénon comprimé à environ 25 atm [9]. Il s'agissait alors de la troisième génération d'appareils TDM. La technologie la plus répandue depuis les années 1990 et jusqu'à aujourd'hui consiste à utiliser de multiples barrettes de cristaux à scintillation couplés à des photodiodes. Ainsi, il est possible d'acquérir plusieurs coupes en même temps. Dans la même génération s'inscrivent les appareils dits à faisceau conique, qui utilisent généralement un détecteur à panneau plat. Une salle de TDM à faisceau conique moderne est montrée à la Fig. 1.1.

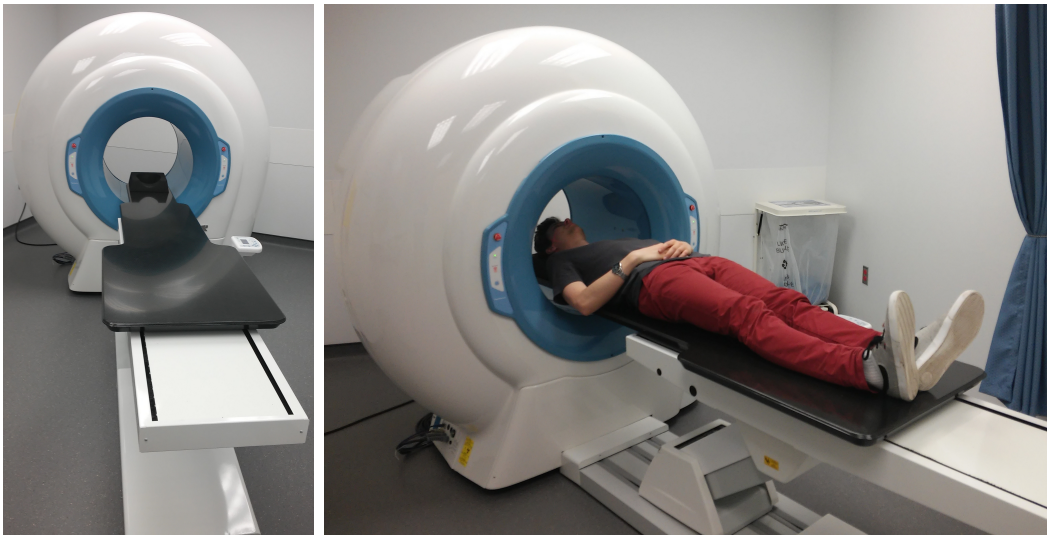


FIGURE 1.1 – Salle de tomodensitométrie à faisceau conique dédiée aux examens de la région dento-maxillo-faciale, modèle 5G produit par NewTom (Vérone, Italie).

1.1.1 Interactions rayonnement-matière en TDM

Il est important de faire un bref survol de la physique des rayons X, ce qui permettra de mieux comprendre certaines propriétés des images en TDM. Le choix de l'énergie du faisceau de radiation est fondé par la physique des interactions entre les tissus humains et les rayons X. En deçà de 30 keV, la plupart des rayons X sont absorbés dans le corps humain sur une distance faible (de l'ordre du cm) et ne peuvent atteindre le détecteur pour former une image. L'intervalle d'intérêt de 30 keV à 140 keV est caractérisé par le meilleur contraste entre les

différents tissus dans les images radiographiques [11]. Le contraste provient de la différence entre les coefficients d'atténuation linéaires μ de différents tissus. Le coefficient μ est le produit de deux quantités : la masse volumique ρ du tissu et le coefficient d'absorption massique μ/ρ. La masse volumique des tissus varie naturellement selon le site anatomique et le coefficient μ/ρ dépend des proportions d'atomes différents qui composent les tissus. Les tissus mous, tels les muscles, s'apparentent à l'eau et sont surtout constitués d'atomes de faible numéro atomique Z, comme l'hydrogène ($Z = 1$) et l'oxygène ($Z = 8$). Ils sont relativement peu atténuants à ces énergies et causent principalement la diffusion Compton des rayons X incidents. Par contre, les tissus durs sont riches en minéraux et contiennent du calcium ($Z = 20$) et du phosphore ($Z = 15$). Ces derniers sont très atténuants et causent principalement l'absorption de rayons X via l'effet photoélectrique. En conséquence, les coefficients d'atténuation massiques sont très distincts pour ces différents tissus. Au delà d'environ 140 keV, deux effets sont observables : (i) la radiation devient plus pénétrante et (ii) l'interaction Compton devient dominante dans tous les tissus. Ainsi, les coefficients d'atténuation massiques des différents tissus deviennent de plus en plus faibles et semblables, ce qui réduit le contraste. Enfin, à nombre de photons constant, la dose de radiation au patient augmente avec l'énergie du faisceau et constitue donc un autre facteur limitant.

1.1.2 TDM à faisceau conique

Il est important de préciser que différents types d'équipement existent pour l'imagerie TDM. Les appareils diagnostiques standard, parfois dits *multibarrettes*, possèdent un champ de vue de grand diamètre et servent à l'imagerie diagnostique générale de tous les sites anatomiques. L'appellation multibarrette vient de la conception du détecteur, composé de plusieurs modules disposés en forme d'arc, qui permettent d'imager plusieurs tranches du patient simultanément. Toutefois, la couverture axiale est modeste, de l'ordre de quelques cm. Ainsi, pour les examens typiques, la table avec le patient doit être déplacée et l'ensemble tube-détecteur doit faire plusieurs rotations. Les appareils de TDM à faisceau conique ou CBCT (cone beam CT) sont conçus pour des examens radiologiques particuliers et sont généralement moins encombrants. La principale différence entre la TDM multibarrette et le CBCT consiste dans la conception du panneau de détection et la collimation du faisceau de radiation. Le CBCT utilise un panneau plat de grandes dimensions composé d'une couche de cristaux à scintillation et d'une matrice de photodiodes. Ainsi, le champ de vue du CBCT est couvert sans translation axiale de la source ou du patient et sans rotations multiples. Un champ de vue suffisamment large est atteint via une grande ouverture angulaire du faisceau en direction axiale. Une autre caractéristique notable des appareils CBCT est une contamination importante des projections 2D par la radiation diffusée dans le sujet. Ceci est dû en partie à la grande ouverture du faisceau de radiation et en partie à la simplicité de la conception et de la géométrie du panneau de détection. Celui-ci est planaire et sensible à la radiation incidente ayant des angles d'incidence variés. En conséquence, il est également sensible à la radiation diffusée provenant de

tout le sujet. En contraste, la TDM multi-barrettes utilise divers moyens pour réduire l'angle d'acceptation de la radiation aux rayons qui proviennent du tube à rayons X en ligne droite.

Quelques exemples d'appareils CBCT conçus pour l'imagerie diagnostique sont des appareils mobiles [12], ainsi que les appareils à petit champ pour l'imagerie à haute résolution du complexe dento-maxillo-facial [13]. Enfin, les appareils d'imagerie des petits animaux (micro-CT) utilisent généralement la géométrie du CBCT [14]. Une autre application importante des systèmes CBCT est le suivi de l'anatomie et du positionnement de patients traités par radiothérapie à faisceau externe. Actuellement, ce type de radiothérapie est effectué grâce aux accélérateurs linéaires médicaux [15], aussi appelés *linacs*. La vérification du positionnement est cruciale pour assurer sa conformité par rapport au plan de traitement [16,17]. Les traitements sont généralement fractionnés et administrés sur plusieurs visites, de sorte que l'anatomie du patient peut changer au cours du traitement. L'image TDM du patient placé pour son traitement permet de confirmer que la position et l'anatomie du patient sont acceptables par rapport aux seuils établis. Cette image est acquise par un appareil CBCT embarqué sur l'accélérateur linéaire traitant le patient. Afin de faire coïncider l'isocentre du linac avec l'axe de rotation du CBCT, la source de rayons X et le détecteur sont attachés au statif du linac par des bras mobiles. Ainsi, la vitesse angulaire de l'acquisition est limitée à celle du statif du linac, ce qui résulte en des acquisitions d'une durée d'environ une minute.

La reconstruction itérative constitue une approche prometteuse pour la réduction de dose en CBCT et l'augmentation de la qualité d'image à dose constante [18–24]. Elle peut donc amener un bénéfice net pour le patient de la première ou de la seconde façon. Ce sujet est exploré dans la présente thèse, au chapitre 2.

1.2 Tomographie optique en dosimétrie

Un autre champ d'application dans lequel s'inscrit le présent travail est la tomographie optique en dosimétrie. La tomographie optique consiste à imager des objets partiellement transparents à la lumière visible. Il est aussi important de distinguer ce mode d'imagerie d'autres types d'imagerie portant des noms semblables, par exemple, la tomographie optique diffuse, qui consiste à imager l'anatomie humaine via la diffusion d'un faisceau laser de lumière visible. La tomographie optique est entre autres utile en dosimétrie pour la radiothérapie [25–28]. La TDM optique permet d'obtenir des profils de dose de radiation en 3D via des gels dont les propriétés optiques changent en fonction de la dose de radiation absorbée. En premier lieu, les notions fondamentales de la dosimétrie 3D par gel seront abordés. En second lieu, quelques aspects importants de la tomographie optique à faisceau conique seront discutés dans le contexte de l'application en dosimétrie.

1.2.1 Dosimétrie 3D par gel

La dosimétrie 3D est une méthode d'assurance-qualité et de validation de protocoles cliniques en radiothérapie. Il s'agit d'une méthodologie devenue particulièrement importante à partir des années 2000. À ce moment, la dosimétrie 3D est devenue nécessaire pour assurer la qualité d'une nouvelle technique de traitement en radiothérapie appelée *radiothérapie conformationnelle avec modulation d'intensité* (RCMI) ou, en anglais, IMRT pour *intensity-modulated radiotherapy* [15, 28, 29]. Cette technique crée des distributions de dose de forme complexe, au moyen d'une séquence de plusieurs irradiations à des angles d'incidence différents et des faisceaux de radiation dont la forme est libre (non-rectangulaire) et peut changer dans le temps. La dosimétrie 3D permet de comparer les distributions de dose planifiées et administrées et ainsi d'assurer la sécurité du patient et la conformité du traitement à la prescription médicale.

La dosimétrie 3D consiste à acquérir des cartes tridimensionnelles de la dose de radiation absorbée par un *fantôme* d'assurance-qualité. Le fantôme sert à jouer le rôle du patient lors de mesures dosimétriques. Il est souvent composé d'un matériau dont les propriétés d'absorption de la radiation ionisante sont semblables à celles de divers tissus humains. Les fantômes qui servent à vérifier les propriétés d'un faisceau de radiation sont souvent de forme simple : cubiques, en plaque etc. Les fantômes qui servent à simuler des traitements de patients imitent le corps humain en forme géométrique et en variation des propriétés des tissus, et sont alors dits *anthropomorphiques*. Il est maintenant approprié de mettre en évidence certaines propriétés des fantômes qui sont fonction des besoins cliniques.

L'eau est un bon choix pour représenter des tissus mous humains en raison de son abondance dans ceux-ci. Toutefois, il est souvent peu commode de travailler avec un liquide et de donner une forme complexe à un tel fantôme. Ainsi, différents matériaux solides ayant des propriétés radiologiques proches à celles de l'eau sont utilisés pour simuler les tissus mous dans les fantômes : résine d'isocyanate [30], polyméthacrylate de méthyle (PMMA) [31], résine d'epoxy ou *eau solide* [32] et diverses formulations commerciales. Dans un fantôme sont incorporés des dispositifs de détection, comme des chambres à ionisation, des semi-conducteurs [15] ou des fibres scintillantes [33]. Ces détecteurs sont la plupart conçus pour faire des mesures ponctuelles, tout en possédant une taille non négligeable, ce qui réduit la résolution spatiale du système de détection. De plus, ils sont parfois constitués de matériaux bien distincts des tissus humains, ce qui demande des corrections complexes lors de l'analyse des lectures. Dans la situation où l'on désire échantillonner l'absorption de radiation dans un espace tridimensionnel avec une grande résolution spatiale, sans perturber le flux de la radiation, une solution technique particulièrement commode est de faire en sorte que le fantôme et le détecteur ne fassent qu'un.

La dosimétrie par gel représente une telle solution [28] en unifiant le fantôme et le détecteur. Les gels utilisés sont constitués à 95% d'eau et d'environ 5% de gélatine et d'additifs spéciaux. Les

additifs rendent le dosimètre sensible à la radiation : les régions irradiées changent de propriétés chimiques et physiques. La gélatine ou une substance semblable empêche le déplacement des additifs dans le contenant et l'altération de la distribution de dose de radiation absorbée. Plusieurs formulations différentes ont été proposées à ce jour, dont quelques-unes qui ont connu un usage plus répandu. Par exemple, la solution de Fricke est historiquement l'un des premiers dosimètres à gel et a subi plusieurs améliorations [34]. Il s'agit d'une solution aqueuse acide et oxygénée d'ions ferreux (Fe^{2+}). Lorsque le dosimètre est irradié, l'eau subit une radiolyse, de sorte à fournir un radical H•. Suite à une chaîne de réactions, l'ion ferreux se transforme en ion ferrique (Fe^{3+}). En conséquence, les régions irradiées changent de couleur, qui passe d'un jaune pâle à un mauve foncé, ce qui en fait un dosimètre dit *radiochromique*. Il s'agit d'une propriété importante, car l'irradiation fait varier la densité optique du gel sans augmenter sa turbidité. Un développement plus récent est le gel de sulfate ferreux combiné avec l'orange de xylenol, aussi connu sous l'acronyme FXG. D'autres gels sont basés sur la réaction de polymérisation de monomères [35] dans une solution aqueuse. Les types de monomères utilisés sont différents selon la formulation, par exemple l'acrylamide et le vinyle. Lors de l'irradiation, les radicaux libres résultant de la radiolyse de l'eau provoquent l'agrégation de monomères en chaînes, ainsi que la réticulation de ces dernières, c'est-à-dire la formation d'une grille tridimensionnelle. En général, la polymérisation provoque l'augmentation de la turbidité du gel, c'est-à-dire sa capacité de diffuser la lumière. Ce type d'interaction porte le nom de la diffusion de Mie, la diffusion d'ondes électromagnétiques planes par des sphères constituées d'un matériau diélectrique dont la taille est comparable à la longueur d'onde de la radiation incidente.

Les dosimètres à gel sont généralement contenus dans des pots transparents de forme cylindrique. Ceux-ci sont insérés dans un fantôme plus grand lors de l'irradiation. Après celle-ci, le changement des propriétés du dosimètre doit être *lu*, c'est-à-dire imagé, de manière non-destructive. Globalement, un gel à base d'ions ferreux est stable du point de vue chimique et physique pendant environ 24 h après l'irradiation, ce qui laisse un temps limité à l'imagerie [36]. Il est donc important de passer à une brève comparaison des méthodes d'imagerie de gels radiosensibles.

Différentes méthodes pour la lecture de gels radiosensibles ont été proposés. Une méthode reconnue pour sa qualité d'image est l'imagerie par résonance magnétique (IRM) [27]. Ce mode d'imagerie identifie la composition chimique des régions du dosimètre via la modification de la direction de précession des spins des atomes du gel [9]. Des gels basés sur l'ion ferreux et les gels à polymérisation peuvent être imagés en utilisant des protocoles d'acquisition distincts [34,35]. Cette modalité est relativement coûteuse et sa disponibilité est variable en clinique. Une alternative récente est la tomographie optique. Il existe différents types de tomographes optiques [27,37,38], où le faisceau lumineux peut être mince, large et parallèle ou divergent (conique). Selon la source de lumière et le type de faisceau, différents agencements

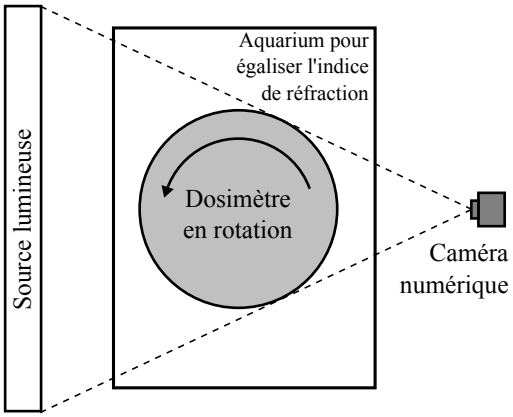


FIGURE 1.2 – Schéma simplifié d'un tomographe optique à faisceau conique.

de miroirs, lentilles et lecteurs optiques sont utilisés. Dans ce travail, il sera question d'appareils à faisceau lumineux conique. Ces derniers sont simples dans leur conception mécanique, peu chers et sont d'entretien facile. Ces appareils sont en principe semblables aux appareils de TDM à rayons X, mais les phénomènes optiques propres à la lumière visible entraînent de nouveaux défis dans la modélisation et la réduction d'erreurs de reconstruction. Ces particularités seront discutées de façon détaillée.

1.2.2 Tomographie optique à faisceau conique

Les appareils de tomographie optique à faisceau conique [39–41] ont pour but de créer des ensembles de projections 2D contenant des mesures d'atténuation partielle de rayons de lumière visible. Ils possèdent une conception semblable à celle de la TDM à rayons X, mais quelques différences sont à noter, voir Fig. 3.1 . L'ensemble source-détecteur est statique et le dosimètre se fixe sur une tête rotative. La source est un écran plat qui émet une lumière diffuse d'une couleur déterminée, c'est-à-dire ayant un spectre de longueurs d'onde relativement étroit. Le dosimètre est confiné dans un *aquarium* de forme à peu près cubique. Il possède deux fenêtres opposées afin que la lumière traverse le dosimètre. L'aquarium est rempli d'eau ou d'un autre fluide, appelé fluide d'égalisation de l'indice de réfraction ou *refractive index matching fluid* en anglais). L'objectif est de s'assurer que le dosimètre et l'aquarium possèdent environ le même indice de réfraction afin que les rayons de lumière voyagent en ligne droite entre la source et le détecteur. L'indice de réfraction est égalisé pour la longueur d'onde moyenne de la source lumineuse, dont le spectre est étroit, comme mentionné ici-haut. Le détecteur est une caméra vidéo numérique. La caméra possède un point focal approximatif de petite taille, où tous les rayons passent avant d'atteindre le détecteur. D'ici vient l'appellation *faisceau conique*. Ici, le détecteur peut être considéré comme un point et la source de lumière comme un panneau plat. Ainsi, la direction du rayonnement est inversée par rapport à la TDM à rayons X.

Plusieurs effets physiques sont présents dans ce système au-delà de la loi d'atténuation de Beer-Lambert, détaillée dans la section suivante. Dans le cas où un dosimètre à polymérisation est utilisé, la turbidité des zones irradiées augmente. Par conséquent la diffusion de la lumière visible engendre des biais dans les lectures de densité. Compte tenu du modèle d'atténuation utilisé dans les présents travaux, les objets à haute turbidité ont été exclus de cette étude. La recherche présentée ici porte donc sur des gels radiochromiques ou semblables. D'autres phénomènes à considérer sont la réfraction due aux parois du contenant de gel et aux légères différences d'indices de réfraction entre le gel et l'aquarium. Une autre considération est la largeur finie du spectre de la lumière émise. Les diodes électroluminescentes (DELs) possèdent un spectre d'émission de largeur non-négligeable. En combinaison avec la sensibilité variable de la caméra à différentes longueurs d'onde, le signal enregistré ne correspond plus au modèle d'atténuation monochromatique. Ce phénomène influence la quantification absolue des coefficients d'atténuation en dosimétrie par gel et sera discuté en détail au chapitre 3.

1.3 Reconstruction tomographique

La reconstruction tomographique est la transformation mathématique qui permet de calculer une carte 3D de coefficients d'atténuation linéaires μ de l'objet imagé à partir d'un ensemble de projections 2D de lectures d'atténuation partielle d'un rayonnement. Les concepts nécessaires à la compréhension de ce sujet sont issus de différents domaines : le traitement de signal et la théorie de l'échantillonnage, le traitement d'images numériques, l'analyse de Fourier, les statistiques, la solution de problèmes inverses mal posés et l'optimisation de fonctions à plusieurs variables. Cette section a pour but d'introduire le lecteur à ces éléments de façon accessible et cohérente.

1.3.1 Modèle direct de l'atténuation

La reconstruction tomographique s'appuie sur un modèle de l'atténuation de la radiation dans l'objet. La formulation du problème inverse s'appuie donc sur la définition du modèle direct. Les présents travaux se basent sur un modèle assez simple, qui est suffisamment juste pour plusieurs applications de la TDM.

On suppose que les photons sont émis par une source ponctuelle et possèdent tous la même énergie E. Ils suivent des trajectoires rectilignes jusqu'au détecteur et on suppose que l'énergie de chaque photon reste inchangée. Sur chaque parcours source-détecteur, aussi appelé *rayon*, une partie des photons est absorbée selon la loi de Beer-Lambert [42] :

$$I = I_0 \exp\left[-\int_L \mu(E,l)dl\right] \approx I_0 \exp\left[-\sum_{j\in J} l_j \mu_j(E)\right]. \tag{1.1}$$

Elle est présentée ci-haut sous forme continue et avec approximation discrète. I désigne l'intensité de la lecture au détecteur à l'acquisition, I_0 l'intensité de référence (sans patient), L

désigne le parcours du rayon, μ le coefficient d'atténuation linéaire pour une énergie de photon E, l la position et dl la variable d'intégration (distance infinitésimale). Dans le cas discret, le patient est constitué de voxels indexés par j, chacun de μ_j fixe, avec un parcours sur des index $j \in J$ et une longueur d'intersection rayon-voxel l_j.

Il est important de mentionner les limites de ce modèle. Les appareils de TDM médicaux sont munis de tubes à rayons X et non de sources nucléaires, car ces tubes permettent de régler l'énergie et le débit de radiation facilement pour s'ajuster au patient et au type d'examen et requièrent moins de précautions de radioprotection. Un tube à rayons X émet un large spectre d'énergies de photons via l'effet de radiation de freinage d'électrons dans une cible en tungstène. La radiation est filtrée par une lamelle métallique ou une combinaison de telles lamelles, de sorte à ce que l'énergie minimale du spectre diagnostique soit d'environ 30 keV. Les matériaux typiques en imagerie diagnostique sont l'aluminium et le cuivre. L'énergie maximale est fixée pour chaque acquisition selon la taille du patient et le site anatomique, à une valeur entre 90 keV et 140 keV. Le coefficient d'atténuation de chaque type de tissu imagé possède sa propre dépendance énergétique $\mu(E)$, qui n'est pas linéaire. Il serait en théorie possible d'estimer $\overline{\mu}$ pour un spectre donné via une moyenne pondérée, mais cette information est peu utile. En fait, le spectre de la radiation change au fur et à mesure que le faisceau traverse le patient et ce spectre *durci* est généralement moins atténué par les tissus [9]. En conséquence, les images reconstruites sous l'hypothèse monochromatique possèdent des densités plus faibles autour du centre de l'image. Cet effet est plus prononcé dans les régions comprises entre deux ou plusieurs os ou implants métalliques, qui modifient dramatiquement le spectre via absorption photoélectrique des photons de basse énergie.

L'hypothèse de la propagation linéaire de la radiation dans l'objet imagé est problématique aussi. La diffusion Rayleigh cause un changement de direction des photons incidents sans affecter leur énergie. La diffusion Compton cause un changement de direction et une réduction de l'énergie de photons diffusés. Ainsi, les lectures d'intensité au détecteur sont biaisées via la détection de photons diffusés, avec une intensité plus élevée au centre du détecteur [9, 43, 44]. Ainsi, les images reconstruites sous l'hypothèse de propagation linéaire ont des densités plus faibles autour du centre de l'image, ce qui amplifie le biais introduit par le durcissement de faisceau. La correction des effets de durcissement de faisceau et de la diffusion dépasse le cadre de cette thèse ; toutefois, la discussion des résultats tiendra compte de ces effets.

Il est important de prendre note des unités typiques pour les variables physiques. En TDM à rayons X, l'énergie des photons est exprimée en kiloelectronvolts (keV). Pour un faisceau polychromatique dans les applications cliniques, on utilise souvent le potentiel du tube radiogène ou kVp (de l'anglais *kilovoltage peak*) qui influence en grande partie les propriétés du fasiceau. En tomographie optique, la dépendance en énergie est remplacée par une dépendance en longueur d'onde λ de la lumière incidente, exprimée en nanomètres (nm). Les distances parcourues sont généralement en centimètres (cm), vu la taille des sites anatomiques ou des

objets à imager. Par conséquent, les coefficients d'atténuation linéaires μ ont des unités de cm^{-1}. En imagerie médicale clinique, le coefficient d'atténuation linéaire n'est pas employé. L'image du patient est plutôt une carte de densités apparentes mesurée en unités Hounsfield ou HU [9], qui est une échelle conventionnelle où l'air possède une densité de -1000 HU et l'eau 0 HU :

$$\mathrm{HU}(\mu) = 1000 \times \frac{\mu - \mu_{\mathrm{eau}}}{\mu_{\mathrm{eau}}}. \tag{1.2}$$

Le maximum de l'échelle est fixé à 3095, afin de stocker la densité comme un nombre entier non-signé de 12 bits. En général, les appareils de TDM cliniques sont étalonnés pour représenter les densités apparentes correctement, via le scan TDM d'un fantôme contenant plusieurs matériaux connus. On retrouve ces matériaux sur les images TDM reconstruites sans étalonnage pour relever leur densité, et une courbe d'étalonnage est établie à partir de densités reconstruites et attendues. Cette démarche peut être entravée par les biais dus au durcissement de faisceau et la radiation diffusée, mais le détail des méthodes de correction dépasse le cadre de cette thèse.

1.3.2 Reconstruction analytique

La reconstruction analytique constitue l'approche la plus répandue en clinique. Cette méthode se base sur le *théorème de la tranche centrale de Fourier* [45]. Soit une image μ d'une tranche d'un patient, qui est une matrice 2D de coefficients d'atténuation linéaires du patient, et M sa transformée de Fourier 2D ou le spectre de l'image. Soit $p(\phi)$ une projection acquise avec un faisceau à rayons parallèles, où ϕ désigne l'angle d'orientation du profil p par rapport au système de coordonnées de l'image. Le théorème stipule que la transformée de Fourier 1D de cette projection $P(\phi)$ se trouve exactement sur une ligne orientée à ϕ dans le plan de M. Ainsi, en théorie, il est possible de reconstruire M à partir de l'ensemble des $P(\phi)$ obtenues à partir des projections acquises. Une transformée de Fourier inverse permet d'obtenir μ :

$$\{p(\phi)\} \xrightarrow{\text{TF 1D}} \{P(\phi)\} \rightarrow M \xrightarrow{\text{TF}^{-1}\ \text{2D}} \mu. \tag{1.3}$$

Le remplissage de l'espace de Fourier de l'image est illustré à la Fig. 1.3. L'approche par transformées de Fourier successives est difficile à utiliser en clinique étant donné que l'acquisition se fait principalement en mode *faisceau en éventail*, ce qui requiert une réorganisation des données vers la géométrie du faisceau à rayons parallèles. De plus, l'espace de M rempli avec l'ensemble $\{P(\phi)\}$ manque d'échantillons en direction angulaire. Ceci nécessite une interpolation sur une grille cartésienne dans le domaine de Fourier, entraînant des artéfacts.

Une approche alternative est la rétroprojection filtrée. À partir du théorème ci-haut, on peut trouver une relation directe entre $p(\phi)$ et μ [42]. Le développement montre que chaque projection peut être « étalée » sur l'image finale et la suivante sera sommée sur l'estimé courant de l'image. Toutefois, un filtre passe-haut est nécessaire pour satisfaire le théorème. Dans l'espace de Fourier 2D, il s'agit d'une *rampe* $|\varrho|$ où ϱ est la fréquence spatiale (en coordonnées

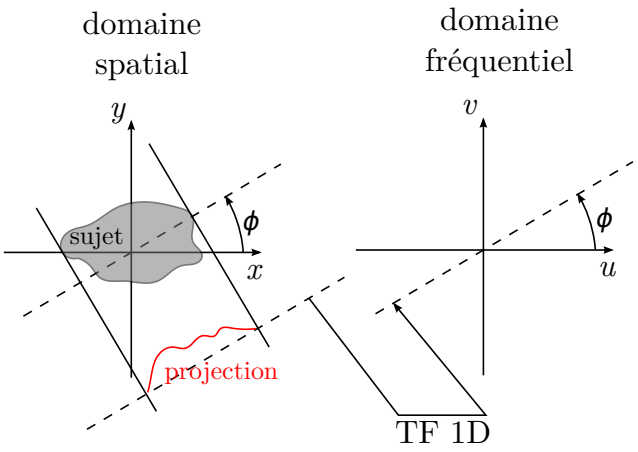


FIGURE 1.3 – Remplissage de l'espace de Fourier 2D d'une tranche de l'image avec les transformées de Fourier 1D des projections.

polaires). Ce filtre ne possède pas de transformée inverse, car son intégrale n'est pas bornée, mais plusieurs approximations sont disponibles [46]. En général, le profil du filtre est amené à 0 pour des fréquences spatiales très élevées, ce qui a pour effet positif de lisser le bruit intrinsèque aux images de TDM. Le principe de la rétroprojection filtrée à été étendu par Feldkamp, Davis et Kress pour être appliqué aux acquisitions avec des appareils multibarettes ou à faisceau conique. L'algorithme, nommé FDK [4], effectue les interpolations nécessaires pour émuler plusieurs acquisitions en faisceau en éventail. Les principaux avantages de cet algorithme sont une complexité du calcul relativement faible et la capacité d'utiliser les données au fur et à mesure qu'elles deviennent disponibles. Il constitue la méthode la plus employée par les fabricants d'appareils de TDM d'imagerie diagnostique, les imageurs embarqués en radiothérapie et les appareils de tomographie optique. Néanmoins, cette méthode est surpassée par différentes méthodes itératives qui formulent le problème de reconstruction comme un problème d'optimisation sous contraintes. Ces dernières seront revues en détail ici-bas.

1.3.3 Reconstruction itérative

La reconstruction itérative (RI) est une alternative très prometteuse à la reconstruction analytique en raison de nombreux avantages potentiels. Il existe de nombreux algorithmes de RI, qui sont distincts selon leur formulation mathématique et leur implantation numérique. La reconstruction itérative constitue le principal sujet récurrent dans la présente thèse. La RI est connue depuis le premier prototype de Hounsfield, qui utilisait la reconstruction dite algébrique, dont il sera question plus bas. La RI en TDM a été confinée à la recherche théorique en raison des besoins en puissance de calcul démesurés pour les ordinateurs de l'époque. Toutefois, la puissance des ordinateurs plus récents a causé un accroissement sans précédent

de l'intérêt pour la RI au début des années 2000 et jusqu'à aujourd'hui [6, 18, 47].

Définition statistique de la reconstruction TDM

Le problème de la reconstruction en TDM peut être défini comme un problème d'estimation de paramètres et résolu par des méthodes statistiques. L'image du patient, c'est-à-dire μ, la carte des coefficients d'atténuation linéaires, est exprimée comme un ensemble de paramètres d'une distribution statistique multivariée. Afin de simplifier la notation, μ est traité comme un vecteur-colonne où sont placés tous les coefficients. Il existe une matrice de projection $\mathbf{A}$, ou matrice-système, qui modélise le passage des rayons X à travers le patient comme une transformation linéaire et incorpore plus ou moins précisément la géométrie et la physique du problème, selon la justesse et précision désirées. En pratique, la matrice-système constitue une approximation assez grossière de la réalité. Le produit matrice-vecteur $\mathbf{A}\mu$ résulte en un vecteur-colonne contenant les mesures aux détecteurs obtenues selon ce modèle simplifié.

De plus, le processus d'acquisition de données est soumis à différents types d'erreurs, comme le bruit quantique lié à la natures aléatoire des interactions rayonnement-matière et le bruit dû aux composantes électroniques du détecteur. Ainsi, des erreurs s'ajoutent au signal. Le modèle le plus simple et le plus souvent employé est celui d'un bruit additif de nature aléatoire, désigné par le vecteur b, ayant une réalisation du bruit par lecture de signal. Dans le contexte statistique, le mot *réalisation* désigne des valeurs concrètes produites au hasard en respectant la loi de probabilité de la variable aléatoire en question. Le choix de la distribution de probabilité du bruit varie selon la complexité du modèle. En somme, le modèle de la projection ou *problème direct* s'écrit :

$$y = \mathbf{A}\mu + b, \tag{1.4}$$

où y désigne le vecteur des lectures aux détecteurs et $\mathbf{A} \in \mathbb{R}^{p \times q}$ où p est le nombre de lectures aux détecteurs et q est le nombre de voxels dans le volume d'intérêt. Compte tenu du fait que chaque voxel est lié à un nombre limité de lectures de détecteurs, à cause de la géométrie d'acquisition, $\mathbf{A}$ est creuse (possède une très grande proportion de valeurs nulles) et est généralement stockée sous forme compressée. En résumé, dans l'équation 1.4, y est connue via l'acquisition TDM, $\mathbf{A}$ est imposée par la géométrie et la physique simulée du système, et μ est le vecteur d'inconnues. Afin de trouver une estimation satisfaisante de μ, il s'agit de (i) bien poser et (ii) résoudre le problème inverse. Les stratégies respectives seront discutées dans les sections suivantes.

Fonction-objectif

En s'appuyant sur le problème direct, on résout généralement le problème inverse en définissant une *fonction-objectif* scalaire multivariée $F(\mu, y)$ dans laquelle μ représente l'ensemble de paramètres libres et le résultat de F est un scalaire. Ce scalaire représente une mesure de la

qualité de l'estimation, de sorte à ce que la minimisation ou la maximisation de F par rapport au vecteur μ pointe sur le meilleur estimé de l'image 3D. Il existe plusieurs types de fonction-objectif, et quelques formes utiles en reconstruction itérative pour la TDM seront discutées ici-bas.

Un des facteurs qui déterminent la forme de F est l'hypothèse sur le type de signal au détecteur. En RI, on suppose généralement que chaque lecture au détecteur y_i est la réalisation d'une variable aléatoire. La nature aléatoire du nombre de photons atteignant le détecteur est assujettie à une loi de Poisson, qui est particulièrement appropriée à basse dose. L'écart-type est alors : $\sigma = \sqrt{N}$ ou N est le nombre de photons incidents sur le détecteur. À dose de radiation standard, c'est-à-dire aux flux de photons élevés, la loi normale de moyenne N et écart type $\sqrt{N}$ constitue une bonne approximation pour représenter le nombre de photons atteignant le détecteur [9].

Estimateurs

Un autre facteur qui détermine la forme de F est le type d'*estimateur* sélectionné pour estimer les paramètres libres μ. Les principaux types d'estimateurs en ordre de complexité croissant sont :

— l'estimateur des moindres carrés (MCar),

— l'estimateur maximum de vraisemblance (MV),

— l'estimateur maximum *a posteriori* (MAP).

Chaque estimateur comporte différentes hypothèses *a priori* distinctes sur les propriétés statistiques des variables à estimer et des mesures expérimentales. Comme on verra plus bas, chaque estimateur plus simple constitue un cas particulier de l'estimateur plus complexe. On va s'affranchir temporairement du contexte de la TDM et emprunter une nomenclature générique rencontrée dans la littérature consacrée à la statistique : soit X la variable aléatoire désignant les mesures expérimentales et x – sa réalisation ; soit Θ la variable aléatoire correspondant à la quantité recherchée et θ la quantité recherchée.

L'estimateur MCar suppose très peu de connaissance sur le système étudié : on pose simplement que les incertitudes sur les mesures sont de distribution normale, sont statistiquement non-corrélées et de moyenne nulle. Ceci en fait un estimateur simple, mais dont la qualité des résultats dégrade rapidement dans des conditions défavorables, c'est-à-dire lorsque les données expérimentales sont peu informatives ou simplement insuffisantes. Cet estimateur consiste à minimiser la norme euclidienne $||.||_{l_2}$ du vecteur-différence entre les lectures expérimentales x et l'estimation du signal $\mathbf{A}\theta$:

$$\theta^* = \arg\min_{\theta} F(x, \theta) = \arg\min_{\theta} ||x - \mathbf{A}\theta||_{l_2}, \tag{1.5}$$

où θ^* désigne la meilleure estimation de l'information recherchée. Il est important de noter que la « meilleure estimation » l'est par rapport aux modèle direct et estimateur choisi, et peut

se révéler insatisfaisante pour l'observateur humain. Cette fonction-objectif est appropriée pour des acquisitions tomographiques à dose standard, dont les données de projection sont peu bruitées. La reconstruction algébrique [46, 48–50] ou ART pour *algebraic reconstruction technique* est un exemple classique d'algorithme utilisant cette fonction-objectif en TDM.

L'estimateur maximum de vraisemblance incorpore plus d'hypothèses *a priori* que la définition précédente. Cet estimateur suppose que les mesures expérimentales suivent une distribution statistique particulière f, qui n'est pas nécessairement une loi normale. Toutefois, la variable à estimer n'est pas contrainte par une connaissance *a priori* ; ainsi, sa fonction de densité de probabilité (FDP) est représentée par une constante sur un très grand domaine. L'estimateur MV s'écrit comme :

$$\theta^* = \arg\max_{\theta} \left[f_{X|\Theta=\theta}(x) \right]. \tag{1.6}$$

Le formalisme peut être étendu à une fonction-objectif multivariée. Pour un vecteur de mesures x, sous les hypothèses que tous les éléments de x suivent la même distribution et constituent des réalisations indépendantes les unes des autres, la vraisemblance est définie comme le produit des FDP f paramétrées par le vecteur de paramètres θ :

$$L_{\text{prod}}(X|\Theta = \theta) = \prod_{i=1}^{n} f_{X|\Theta=\theta}(x_i). \tag{1.7}$$

Cette fonction possède une propriété remarquable : étant donné un vecteur d'échantillons x, la vraisemblance est maximale lorsque θ correspond aux paramètres les « plus vraisemblables » ayant servi à générer l'échantillon x obéissant à la loi f. En pratique, on s'intéresse à la vraisemblance logarithmique :

$$L(X|\Theta = \theta) = \ln(L_{\text{prod}}) = \sum_{i=1}^{n} \ln f_{X|\Theta=\theta}(x_i). \tag{1.8}$$

Étant donné que le logarithme est une fonction strictement croissante, l'argument qui maximise L_{prod} et L est le même. Comme les méthodes d'optimisation s'intéressent à la dérivée de L, cette dernière est souvent plus commode à dériver terme à terme que L_{prod}. L'estimateur MV s'écrit comme suit :

$$\theta^* = \arg\max_{\theta} L(X|\Theta = \theta). \tag{1.9}$$

Sous les hypothèses énoncées pour l'estimateur MCar, l'estimateur MV se réduit à l'estimateur MCar. Quelques algorithmes de reconstruction bien connus qui utilisent le maximum de vraisemblance comme fonction-objectif sont l'algorithme de la maximisation de la vraisemblance par espérance-maximisation (MLEM), de l'anglais *maximum likelihood expectation-maximization* [51], l'algorithme dit *convexe* [52, 53] et *convexe à sous-ensembles ordonnés* (OSC) [54], de l'anglais *ordered subsets convex*. Ce dernier sera abordé en grand détail dans la présente thèse.

L'estimateur maximum *a posteriori* incorpore une hypothèse supplémentaire par rapport à l'estimateur précédent : on suppose qu'on connaît la distribution statistique de la variable à estimer, qui n'est pas nécessairement une distribution uniforme. On note la FDP de la variable estimée θ comme f_θ. Cette fonction permet de contraindre les valeurs possibles de θ dans un domaine considéré raisonnable et ainsi injecter de l'information supplémentaire dans un modèle autrement sous-déterminé. Par conséquent, cet estimateur est approprié pour les situations où les données expérimentales sont incomplètes et/ou le bruit est fort. L'estimateur MAP prend une forme commode à utiliser en pratique grâce au théorème de Bayes. Pour simplifier la dérivation, supposons que l'on tente d'estimer une quantité scalaire θ. La FDP de la variable aléatoire correspondante, connaissant la réalisation x, s'écrit comme :

$$f_{\Theta|X=x}(\theta) = \frac{f_{\Theta X}(\theta, x)}{f_X(x)} \tag{1.10}$$

$$= \frac{f_{X|\Theta=\theta}(x) f_\Theta(\theta)}{f_X(x)}. \tag{1.11}$$

Comme on cherche à maximiser cette fonction par rapport à θ, le dénominateur constitue une simple constante de normalisation. L'estimateur MAP résultant s'écrit comme :

$$\theta^* = \arg\max_\theta \left[f_{X|\Theta=\theta}(x) f_\Theta(\theta) \right]. \tag{1.12}$$

Cette fonction-objectif peut être étendue au cas multivarié de manière semblable à celle montrée pour l'estimateur MV. Sous l'hypothèse spécifique de l'estimateur MV que $f_\Theta(\theta) = \text{const}$, l'expression se réduit automatiquement à l'équation (1.6). Dans les formulations pratiques, on s'intéresse au logarithme de la fonction-objectif, et l'estimateur MAP s'écrit comme une fonction-objectif à deux termes de la forme suivante :

$$F(x, \theta) = L(x, \theta) + \lambda P(\theta), \tag{1.13}$$

où L est le terme de correspondance aux données, semblable à celui de l'estimateur MV, P est un terme de pénalisation qui exprime la contrainte sur θ et λ est un coefficient empirique qui contrôle l'influence de la contrainte sur la solution recherchée. Dans le contexte de la RI en TDM, le terme de pénalisation assure généralement la *régularisation* de l'image, c'est-à-dire la réduction du bruit inhérent au processus d'imagerie. L'estimateur MAP est fréquemment utilisé dans ce contexte en utilisant diverses formulations du terme de pénalisation [52, 53, 55–60]. Cette approche sera également utilisée dans cette thèse ; toutefois, il est important de remarquer qu'il s'agira d'emprunter le concept général de l'estimation MAP plutôt que d'utiliser sa définition formelle dans l'algorithme développé.

La définition de la fonction-objectif, tout dépendant des hypothèses, peut donner une formulation plus ou moins commode à optimiser. Il est désirable d'arriver à une fonction convexe (ayant un maximum unique) et ayant une solution analytique. Ceci est plutôt rare et les méthodes d'optimisation sont souvent conçues pour accommoder une fonction-objectif particulière.

Optimisation de la fonction-objectif

Une fois que la fonction-objectif est posée, il existe plusieurs méthodes d'optimisation permettant de minimiser ou maximiser cette fonction-objectif. Leurs propriétés les plus importantes sont la garantie de la convergence vers l'optimum et la vitesse de convergence, c'est-à-dire le nombre d'itérations nécessaires pour estimer la position de l'optimum de la fonction-objectif avec une tolérance donnée. Elles diffèrent par leur sophistication mathématique, leur complexité numérique et les besoins en mémoire d'ordinateur.

Dans le contexte de la TDM, chaque itération d'une méthode d'optimisation a pour effet d'apporter une correction à l'image 3D μ, dans le but de se rapprocher de l'optimum de la fonction-objectif. Du point de vue numérique, il s'agit d'apporter des correctifs aux valeurs courantes des voxels, en fonction des données de projection expérimentales et estimées par projection directe dans l'image estimée. En conséquence, il est important de retenir que chaque itération d'un algorithme itératif constitue une séquence typique :

— projection à travers l'estimation μ pour obtenir les projections 2D estimées,

— calcul et propagation des facteurs de correction dans l'image 3D (rétroprojection).

Selon la définition de l'algorithme, le calcul des facteurs de correction peut constituer une étape à part entière ou s'intégrer à l'étape de rétroprojection. Une autre caractéristique de l'algorithme de reconstruction est l'estimation initiale μ_0, qui peut être une constante ou le résultat d'une reconstruction analytique. Dans cette thèse, μ_0 est une image uniforme, c'est-à-dire de coefficient d'atténuation constant sur tout le volume.

Une classe importante de méthodes d'optimisation porte le nom de méthodes *du gradient*. Ces dernières calculent le gradient de la fonction-objectif afin de déterminer une direction de changement de l'image 3D estimée qui pointe vers le maximum de la fonction- objectif [61].

La présente thèse utilise une approche très semblable à une méthode du gradient sans y être identique, appelée espérance-maximisation [62]. En termes généraux, à chaque itération d'index n, cette méthode propose une fonction de substitution de la fonction-objectif, la première étant plus simple à maximiser numériquement. Cette fonction de substitution est de forme convexe et est tangente à la fonction-objectif au point de l'image estimée courante $\mu^{(n)}$. De plus, le maximum de la fonction de substitution est localisé à un point $\mu^{(n+1)}$, plus proche de la position du maximum de la fonction-objectif. À chaque itération, la fonction de substitution proposée est maximisée, résultant en un rapprochement vers le maximum global. D'autres exemples de méthodes bien connues pour l'optimisation de fonctions multivariées sont la méthode de descente de coordonnées itérative [63] (ICD), de *iterative coordinate descent* en anglais, ainsi que des variantes de la méthode de Broyden–Fletcher–Goldfarb–Shanno (BFGS) [64, 65].

Une considération importante pour les algorithmes itératifs est ce qu'on nomme le *critère*

d'arrêt. Il s'agit d'une condition mathématique qui, une fois satisfaite, marque l'arrêt des calculs et l'enregistrement de l'estimation finale de l'image recherchée. Le rapprochement vers un gradient nul est un critère d'arrêt possible pour une fonction objectif strictement convexe (à maximum unique). En pratique, les fonctions-objectif combinées ayant plusieurs maxima locaux, ainsi que la discordance entre le modèle mathématique et la réalité physique du système modélisé empêchent d'utiliser un critère d'arrêt analytique. On se rabat alors sur des critères empiriques qui mesurent la qualité de l'image résultante, voire une analyse visuelle de la séquence d'estimées de l'image 3D. Comme mentionné à quelques reprises dans la présente section, la correspondance entre le modèle mathématique et les processus physiques influence la convergence des algorithmes itératifs. En suivant cette logique, la section suivante sera consacrée à explorer les modèles de la géométrie de l'acquisition de projections en TDM.

1.4 Modélisation de la géométrie d'acquisition

Le modèle de la géométrie de l'acquisition est une composante essentielle de tout algorithme de reconstruction et a un impact direct sur la qualité d'image. Quelques propriétés importantes de ce modèle sont : la représentation discrète de l'image 3D, la représentation des traces des rayons X passant à travers le milieu, la forme de la source de radiation et la représentation des éléments du détecteur de radiation 2D. En général, une représentation plus fidèle de la géométrie en TDM requiert plus de ressources numériques, c'est-à-dire des opérations arithmétiques et/ou quantité de mémoire pour contenir le modèle. Toutefois, certaines approximations judicieuses peuvent amener un bon équilibre entre les exigences en ressources informatiques et la justesse du modèle.

1.4.1 Représentation de l'image 3D

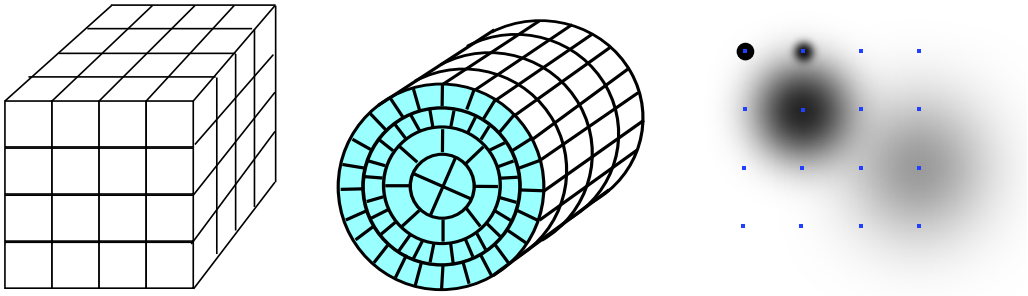
La représentation de l'image 3D relève des concepts fondamentaux d'échantillonage de signaux et images. Un choix très commun est de supposer que le champ de vue de l'appareil est un parallélépipède droit (bloc) divisé en *voxels* de petite taille et ayant également la forme de blocs, voir Fig. 1.4 (a). En général, la taille des voxels est uniforme dans tout le champ de vue. Pour certaines applications, le champ de vue peut posséder une région d'intérêt constituée de voxels de plus petite taille [66]. Une autre approche est de poser une grille discrète inspirée des coordonnées cylindriques [67–69]. Le champ de vue est alors un cylindre à base circulaire. Il est divisé en tranches adjacentes de forme cylindrique, chaque tranche est divisée en anneaux, et chaque anneau, en sections d'intervalle angulaire uniforme, voir Fig. 1.4 (b). Afin d'uniformiser le plus possible la taille des voxels, le nombre de sections par anneau augmente avec le rayon de l'anneau. Cette représentation est surtout intéressante pour réduire la complexité numérique des opérateurs de projection et rétroprojection et de réduire le temps de reconstruction itérative. Du point de vue de l'influence du modèle cylindrique sur les propriétés de convergence des algorithmes itératifs, cette approche pose un nombre de nouveaux

défis uniques qui dépassent le cadre de cette thèse. Une alternative intéressante aux voxels discrets est la représentation des densités par une superposition de fonctions continues [70]. Typiquement, ce modèle se place dans une grille cartésienne en 3D avec un nombre fini de points-repères répartis régulièrement. Sur chaque point est centrée une fonction représentant le coefficient d'atténuation de radiation local, aussi appelée fonction de base ou *blob*, voir Fig. 1.4 (c). Afin de simplifier les calculs associés à la projection et rétroprojection, on privilégie des fonctions possédant une symétrie sphérique et s'annulant sur un domaine fini. En particulier, la fenêtre de Kaiser-Bessel a été étudiée comme fonction de base pour une telle représentation [71, 72]. Les paramètres typiques de la fonction de base sont le coefficient d'atténuation linéaire maximal, le rayon maximal et un coefficient qui contrôle le comportement aux bords (chute abrupte ou graduelle). Ce modèle permet de (i) bien modéliser des frontières continues entre les structures imagées et (ii) atténuer les artéfacts de très haute fréquences spatiale qui se manifestent dans les images aux voxels discrets [72]. En revanche, il est d'une complexité de calcul plus élevée que les méthodes ayant recours à la grille cartésienne et est sujet à une réduction de la résolution spatiale dans certaines conditions [73]. En considérant plusieurs facteurs mentionnés ici-haut, on opte pour une grille cartésienne dans ce travail.

Le modèle à voxels discrets suppose généralement qu'à l'intérieur des frontières du voxel, le coefficient d'atténuation linéaire μ est uniforme. Cette hypothèse peut se révéler problématique si une frontière entre deux structures passe au milieu d'un voxel. Dans ce cas, on observera un effet de *volume partiel*, où l'estimation de μ tend vers une moyenne pondérée des deux μ présents dans le voxel. Du point de vue numérique, l'image 3D peut être stockée de différentes façons. En clinique, on utilise l'échelle Hounsfield (nombres entiers), une représentation discrète suffisamment fine, sachant que le bruit et les différentes causes de biais génèrent des écarts au delà de 1 HU [74]. Toutefois, en RI, un modèle de l'atténuation plus sophistiqué requiert l'utilisation de μ (en $\mathrm{cm}^{-}1$) sur une échelle absolue. Comme ces valeurs sont utilisées comme exposant dans une expression, il est plus prudent de stocker μ avec une précision numérique plus raffinée. Dans ce travail, les nombres à virgule flottante à 32 bits sont utilisés.

1.4.2 Source et détecteur

La source de radiation, comme le point focal d'un tube à rayons X, possède une surface finie. Il s'agit généralement d'un rectangle dont la longueur des côtés est de l'ordre de 0.1 mm. La taille est petite afin d'assurer une bonne résolution spatiale, en réduisant le lissage de l'image dû à la pénombre des structures imagées. La source produit un faisceau divergent, et on peut estimer la position d'un point focal virtuel d'où semblent provenir les rayons X. Il s'agit là d'une approximation, car chaque élément de la source est en soi une source ponctuelle d'un faisceau divergent. Comme pour la majorité des applications cliniques de la TDM, la source de radiation est considérée ponctuelle dans le cas des géométries utilisées dans cette thèse. Pour les simulations entièrement numériques, on assume que l'ouverture du faisceau (collima-



(a) (b) (c)

FIGURE 1.4 – Représentations de l'image utilisées en reconstruction tomographique : (a) Voxels en forme de blocs, (b) représentation cylindrique discrète et (c) fonctions continues à symétrie radiale superposées (montrées en 2D ici). Les modèles (a) et (b) sont de type discret, alors que le modèle (c) est de type continu.

tion) correspond parfaitement à la taille du champ de vue désiré et que le détecteur possède exactement la taille nécessaire. Pour les données expérimentales, on se fie sur les informations du fabricant sur l'appareil et le protocole d'acquisition pour déterminer les mesures liées à la collimation.

Le détecteur de radiation est physiquement une grille de surfaces radiosensibles de forme rectangulaire ou carrée, correspondant aux *pixels* des projections. Certains types de détecteurs possèdent des sillons entre les éléments de détecteur, dont la taille est petite, mais finie. Une autre propriété importante du détecteur est la contamination du signal d'un élément de détecteur par le signal atteignant les détecteurs voisins, appelée *diaphonie* en français et *detector cross-talk* en anglais. Ceci provoque un lissage des projections 2D et une dégradation de la résolution spatiale. La présence de sillons et le lissage de l'image dû à la diaphonie sont rarement pris en charge dans les algorithmes de reconstruction, mais pourraient s'intégrer dans un modèle numérique très fidèle de la projection radiographique. Dans la présente thèse, on considère un détecteur dont les éléments sont rectangulaires, sans sillons et sans interférence entre les lectures de signal.

Un détecteur de radiation réel peut également être muni d'une grille anti-diffusion, constituée de lamelles métalliques orientées de sorte à absorber une partie des rayons X dont la provenance est autre que le point focal de la source. La modélisation de la diffusion dépasse le cadre de cette thèse, on assume donc que le détecteur possède une grille anti-diffusion idéale, qui absorbe toute radiation diffusée.

1.4.3 Intersection entre la radiation et le sujet

Ayant défini les modèles des composantes du système d'imagerie, il est maintenant essentiel de se pencher sur le modèle géométrique de l'intersection entre la radiation et le sujet imagé. Du point de vue historique, les premier algorithmes de reconstruction de type algébrique posaient une hypothèse très simple : sur une droite entre le point focal de la source et le centre d'un élément de détecteur, tous les voxels (de taille finie) traversés étaient considérés comme participant à l'atténuant du rayon, avec des poids relatifs égaux [48]. Cette approche simpliste a été rapidement remplacée par différentes approximations plus fidèles dont il sera question plus bas. Il est important de remarquer que l'objectif de ces méthodes est d'établir un ensemble de voxels J_i traversés par un faisceau de radiation mince ou infiniment mince d'index i. De plus, il s'agit d'attribuer des poids w_{ij} afin de différencier la contribution des différents voxels à l'atténuation du faisceau i. Pour la suite de la thèse, une quantité utile est définie ici-bas :

$$t_i \equiv \sum_{j \in J} w_{ij} \mu_j. \tag{1.14}$$

Elle sera désignée *atténuation totale* ou cumulative sur le parcours i. Il s'agit de l'argument de l'exponentielle dans l'équation 1.1, sous l'approximation d'un faisceau mono-énergétique et d'un sujet discrétisé. La fraction des photons atteignant le détecteur sur ce parcours est alors $\exp(-t_i)$. Cette quantité est principalement utilisée par alléger l'écriture.

Méthodes d'interpolation des poids dans le modèle de l'atténuation

Les méthodes d'interpolation sont les plus populaires pour les applications cliniques de la reconstruction analytique, en raison de leur faible complexité numérique. La méthode la plus répandue est l'interpolation bi-linéaire. Elle est surtout utilisée pour effectuer l'opération de rétroprojection, c'est-à-dire la propagation des facteurs de correction dans le volume, mais peut être reformulée pour la projection directe. Dans ce formalisme, les poids ne sont pas accordés aux voxels traversés par les rayons, mais aux pixels du détecteur. On considère une droite qui relie le point focal de la source et le centre d'un voxel d'intérêt j, voir Fig. 1.5. Cette droite possède aussi un point d'intersection avec le plan du détecteur. On sélectionne les quatre pixels du détecteur les plus proches du point de chute du rayon et on calcule les poids relatifs w_x et w_y. On corrige la valeur du voxel par une moyenne pondérée des facteurs de correction associés aux quatre pixels. Une variation de la méthode consiste à créer une fonction-fenêtre qui pondère les contributions d'un plus grand nombre de pixels [75].

Compte tenu de sa définition, cette approche a tendance à réduire légèrement la résolution spatiale des images 3D, car on échantillonne seulement les lignes de parcours de la radiation au centre de chaque voxel. Du point de vue de l'implantation, cette méthode fait partie de la classe de méthodes dites *guidées par les voxels*, en anglais, *voxel-driven*. Ces méthodes sont très commodes pour effectuer la rétroprojection : pour chaque voxel, on liste les trajectoires

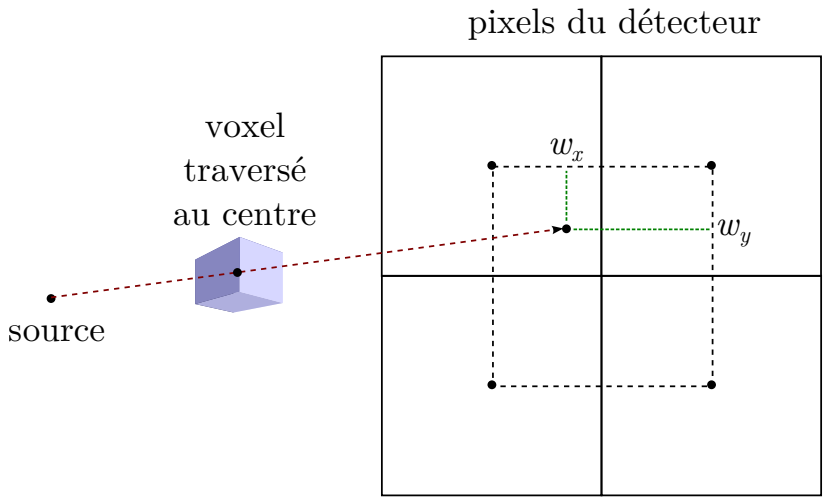


FIGURE 1.5 – Interpolation bilinéaire des contributions des éléments du détecteur pour une ligne d'atténuation déterminée par les positions de la source et du centre du voxel. Les poids relatifs w_x et w_y sont basés sur le point de chute du rayon et sont compris entre 0 et 1. Ils permettent de calculer la moyenne pondérée des contributions des pixels. La figure n'est pas à l'échelle.

source-détecteur et on interpole les facteurs de correction à partir des projections estimées. La projection directe, quant à elle, ne constitue pas une étape des algorithmes analytiques et devient seulement nécessaire pour les algorithmes itératifs. Il existe quelques méthodes très semblables en principe pour effectuer l'interpolation sur les voxels lors d'une projection simulée : voir, par exemple, Joseph [76], où on interpole sur les valeurs de μ de deux voxels voisins traversés par le rayon qui relie la source et le centre du pixel du détecteur, ainsi que l'interpolation par rapport à plusieurs voisins [73]. Ces méthodes sont moins populaires en RI, en raison de la préférence pour des modèles plus représentatifs de la réalité physique du problème.

Des méthodes d'interpolation avancées ont fait leur apparition dans les années 2000, offrant un meilleur équilibre entre temps d'exécution et justesse du modèle. Les méthodes les plus remarquables sont la projection et rétroprojection guidée par la distance [77], ou DD, de *distance-driven projection and backprojection* en anglais, et via empreintes séparables [78] (SF), de *separable footprints* en anglais. Ces méthodes sont de définition assez sophistiquée et ne seront pas discutées en détail dans ce travail. Il est important de retenir que ces méthodes possèdent une justesse variable en fonction de l'angle d'acquisition de la projection ; toutefois, la méthode des empreintes séparables est plus juste que la méthode guidée par la distance, au coût d'une complexité numérique plus élevée [78]. Il est à noter que ces méthodes calculent seulement des poids relatifs et non pas des longueurs de parcours de la radiation, ce qui nécessite des facteurs de conversion lors de la projection directe. On va s'intéresser maintenant

aux méthodes qui calculent les distances physiques.

Modèle aux rayons fins

Une représentation souvent employée pour effectuer les projections directes est le modèle aux rayons fins, et l'algorithme numérique associé est connu sous les nom de *traçage de rayons* et algorithme de Siddon [79]. Ce modèle suppose une droite entre le point focal de la source et le centre de chaque pixel de détecteur, voir Fig. 1.6. Pour chaque droite d'index i, on établit un ensemble de voxels J qui sont traversés par celle-ci, à partir d'une liste d'intersections avec les plans de la grille discrète. Ensuite, pour chaque voxel d'index $j \in J$, on calcule les coordonnées du point d'entrée P_1 et du point de sortie P_2 à partir de la liste d'intersections. La distance de parcours dans le voxel est alors $l_{ij} = |P_1 - P_2|_{ij}$. La méthode a été accélérée de

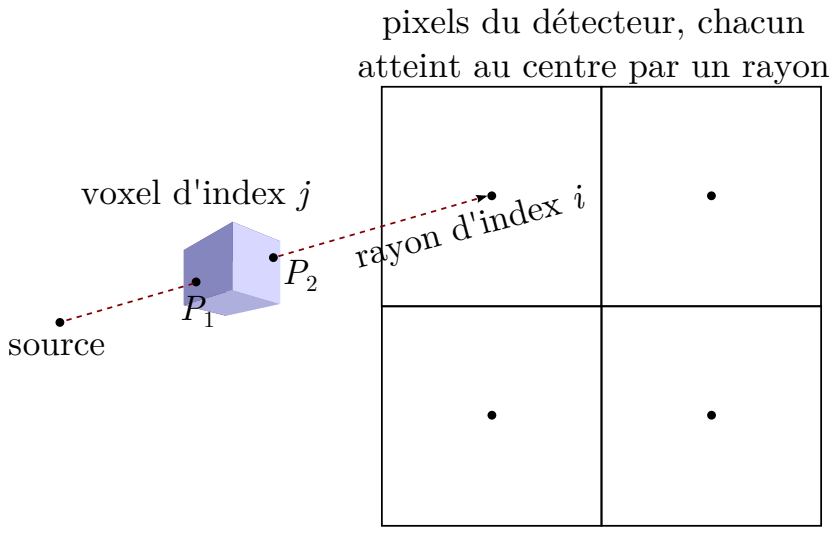


FIGURE 1.6 – Modèle géométrique aux rayons fins. La distance d'intersection entre le rayon d'index i et le voxel d'index j est $l_{ij} = |P_1 - P_2|_{ij}$. La figure n'est pas à l'échelle.

façon algorithmique, portant le nom de l'algorithme de Siddon incrémental [80, 81]. Comme on pose généralement que le coefficient d'atténuation μ_j est uniforme dans le voxel, on peut calculer l'atténuation due à ce voxel sur ce parcours : $l_{ij}\mu_j$. Cette représentation s'intègre parfaitement dans les modèles de l'atténuation exponentielle. Pour cette raison, le modèle est reconnu comme l'un des meilleurs pour simuler les projections directes 2D en reconstruction itérative [73]. Toutefois, lorsque ces longueurs de parcours sont employées pour effectuer la rétroprojection, le sous-échantillonnage du détecteur (dont on considère seulement le point central) dégrade les images 3D avec un artéfact de haute fréquence semblable aux patrons de Moiré, de façon non-négligeable. En pratique, cet artéfact peut être atténué avec une approche MAP incorporant un filtre passe-bas. Compte tenu de l'importance d'un modèle de projection directe fidèle en RI, le modèle aux rayons fins sera utilisé dans la présente thèse.

Modèles alternatifs

Il existe d'autres modèles dérivés du modèle aux rayons fins dont il est important de faire mention. Le modèle à faisceau solide [73, 82, 83], désigné *beam-box model* en anglais, considère que le faisceau de radiation possède la forme d'une pyramide à base rectangulaire (élément du détecteur) et dont le sommet est le point focal de la source. Les poids w_{ij} sont basés sur la partie du volume partiel de la pyramide i contenu dans le voxel j. Ce modèle possède l'avantage de modéliser la taille finie de l'élément de détecteur, assurant une meilleure définition spatiale des contours de l'image. Toutefois, une intersection de volumes fait abstraction de la nature exponentielle de la loi de Beer-Lambert. Ce modèle calcule en quelque sorte un poids moyen issu de plusieurs rayons linéaires, qui sera ensuite inséré dans le modèle de l'atténuation. Pour pallier à ce dernier problème, il est possible d'utiliser le modèle aux rayons fins multiples [73,84]. Ce modèle est une extension logique du modèle aux rayons fins. On sélectionne un nombre fini K de rayons fins reliant la source et le détecteur. La source aura ainsi plusieurs points de départ, basés sur sa taille réelle, et le détecteur aura un nombre de points de chute de la radiation basés sur sa taille réelle. Chaque rayon fin possède un ensemble de voxels traversés J_k. La projection directe utilisant des rayons multiples va respecter la loi de Beer-Lambert, car les éléments de l'intensité du signal seront sommés après son application :

$$I_i = \frac{I_{0,i}}{K} \sum_{k=1}^{K} \exp\left[- \sum_{j \in J_k} l_{ijk} \mu_j \right]. \tag{1.15}$$

Cette approche est de plus en plus juste lorsque le K augmente. En pratique, on utilise souvent 4 rayons par élément de détecteur, en considérant qu'un sur-échantillonage de 2× dans chaque direction du détecteur 2D est suffisant. Le modèle aux rayons fins multiples est utilisé dans la présente thèse pour générer des projections simulées de grande qualité sur un fantôme anthropomorphique. Le modèle aux rayons fins multiples requiert d'importantes ressources matérielles pour effectuer les calculs. Ainsi, cette méthode est peu populaire en clinique, où le temps de calcul est de grande importance. Avec les possibles applications cliniques en tête, les aspects numériques de la modélisation géométrique méritent donc une discussion propre à eux.

1.4.4 Considérations numériques

La reconstruction itérative est caractérisée par une séquence de plusieurs projections et rétroprojections. La taille d'un ensemble de projections est de l'ordre de 500×500 pixels ×500 angles de projection, donc 125 millions de lectures de détecteurs. La taille du volume 3D à reconstruire est du même ordre de grandeur. Avec une précision de 32 bits par élément, on a deux vecteurs d'environ 500 MiB chacun. Dans le but de maximiser la bande passante, c'est-à-dire le débit de données de la mémoire vers le processeur central (CPU) et vice versa, ces vecteurs sont typiquement chargés dans la mémoire vive de l'ordinateur et y résident pendant toute

la reconstruction. Seulement l'estimation finale du volume est écrite sur le disque dur pour stockage permanent. Maintenant, il est important d'estimer la taille de la matrice-système. Comme chaque rayon traverse le volume sur une ligne, il y aura environ 1000 intersections par rayon. Le nombre de poids w_{ij} est donc de l'ordre de 125 milliards. Encore, avec 32 bits par élément, ceci représente environ 500 GiB de mémoire, ce qui dépasse la capacité de la mémoire vive des ordinateurs actuels typiques. Pour les approches d'interpolation, ces poids sont calculés à la volée, lors de la projection et rétroprojection et la complexité de ces opérations est relativement faible. Par contre, le traçage de rayons est une opération assez complexe. En RI, le calcul répétitif des distances l_{ij} est susceptible d'occuper un temps de calcul non-négligeable [85–88]. Ainsi, les méthodes de compression de données, de réutilisation de données pré-calculées et de re-calcul rapide sont de mise pour implanter une reconstruction rapide. Dans le cadre de cette thèse, cette problématique sera discutée en détail dans les chapitres 2 et 4. Avec cet aperçu des défis de l'implantation numérique, il est approprié de discuter des outils de calcul informatique de pointe disponibles pour s'acquitter de la tâche.

1.5 Calcul informatique de pointe et matériel graphique

Comme il a été discuté précédemment, le matériel de calcul numérique est une composante essentielle de la TDM. La reconstruction analytique se fait facilement sur un ordinateur de bureau en environ une minute ou quelques minutes tout au plus ; toutefois, la RI constitue un défi de taille avec une séquence de plusieurs projections et rétroprojections et possiblement un modèle sophistiqué de la géométrie. Pour cette raison, la RI fait appel au calcul informatique de pointe (CIP), une discipline qui étudie l'utilisation optimale de super-ordinateurs et matériel connexe dans le but de résoudre une panoplie de problèmes numériques.

1.5.1 Initiation au CIP

Depuis le début du calcul numérique sur ordinateurs électroniques dans les années 1940, l'utilisateur scientifique ou l'ingénieur a tendance à exiger l'exécution de calculs plus volumineux en moins de temps pour atteindre des objectifs dans la recherche fondamentale, appliquée ou pour rehausser les performances de dispositifs dépendants d'un ordinateur. Les universités, les forces militaires et les firmes d'ingénierie ont donc historiquement été les principaux joueurs dans le développement du CIP. La loi empirique de Moore stipule que le nombre de transistors sur les circuits intégrés augmente de façon exponentielle au fil du temps. Typiquement, on observe que le nombre de transistors double aux deux ans. Toutefois, le désir d'obtenir une performance supérieure en utilisant le matériel informatique « du jour » a motivé la conception des super-ordinateurs, qui effectuent des calculs en parallèle sur plusieurs processeurs [89]. Les architectures de super-ordinateurs ont évolué dans le temps. Dans les années 1970, les machines vectorielles (opérant sur des vecteurs de données plutôt que sur des nombres scalaires) ont été proposées. Dans les années 1980 sont apparus les ordinateurs à processeurs multiples

et mémoire partagée. Ensuite, les années 1990 ont donné naissance à des *grappes de calcul* constituées d'ordinateurs de type plus conventionnel, dont la communication est assurée par une interface de *passage de messages*. L'architecture de type grappe est encore très utilisée de nos jours. Toutefois, la validité des prévisions loi de Moore est remise en question, en raison de la miniaturisation extrême des transistors. En fait, le fonctionnement des composantes à base de semi-conducteurs est lié à leur structure cristalline, et la taille des transistors se rapproche de plus en plus de celle d'un atome, amenant de nouvelles contraintes relevant de la physique quantique. De nouvelles approches de conception de matériel sont en cours de développement pour remplacer les semi-conducteurs conventionnels, par exemple, l'électronique moléculaire, qui promet de miniaturiser davantage les circuits en utilisant des molécules comme composantes logiques [90]. Cette discipline pourrait révolutionner le matériel informatique et, une fois disponible, proposer de nouvelles capacités de calcul et imposer de nouvelles contraintes sur la programmation orientée vers la performance numérique. En attendant la maturation de cette approche, l'utilisation de matériel avec des milliers de processeurs et des interfaces de programmation spécialisées est la voie de l'avenir, du moins à court et moyen terme.

1.5.2 Matériel graphique en CIP

Une innovation récente dans le domaine du CIP est l'introduction de *co-processeurs*. Il s'agit de matériel qui dépend d'un ordinateur classique possédant un processeur central, une mémoire vive et un bus de communication. Comme la vitesse de l'horloge des processeurs arithmétiques et logiques (ALU) d'aujourd'hui plafonne à quelques GHz en raison des propriétés des semi-conducteurs, l'innovation dans le domaine des composantes de base est présentement orientée vers le calcul dit massivement parallèle [91]. Le concept du co-processeur graphique remonte à des efforts de certains scientifiques de convertir le matériel graphique (GPU) en matériel de calcul d'usage général (*general purpose GPU* ou GPGPU en anglais) [47]. Initialement, ces chercheurs ont implanté quelques algorithmes de calcul général sous forme de routines d'affichage afin d'accéder à la puissance de calcul des GPU. Suite à leurs succès, les fabricants d'équipement comme NVIDIA® (Santa-Clara, CA) ont proposé des interfaces de programmation générale conviviaux comme CUDA™ [92] et des GPU compatibles ou carrément dédiés au GPGPU, qui ont pris l'appellation de co-processeurs. Des interfaces standardisées de plus haut niveau ont fait leur apparition ensuite, comme OpenCL, qui permet d'exploiter les GPUs de différents fabricants [93]. En date d'aujourd'hui, le matériel graphique est utilisé par plusieurs librairies de CIP, avec des applications modélisation de processus physiques, des interactions moléculaires, en intelligence artificielle et dans plusieurs autres disciplines [94]. Afin de mieux comprendre les stratégies d'accélération des calculs avec le GPU, il est essentiel de s'initier aux particularités de cette plateforme de calcul.

1.5.3 Conception et performance du GPU

Le GPU, dont un exemple est montré à la Fig. 1.7, est un co-processeur connecté à l'ordinateur via un *bus* parallèle. Il reçoit des données via ce bus, les stocke dans sa propre mémoire

FIGURE 1.7 – Vue externe d'un GPU récent (NVIDIA GTX 1070). Reproduit de https://en.wikipedia.org avec permission.

vive et effectue des calculs sur des dizaines de multi-processeurs. Chaque multi-processeur est muni d'une centaine de coeurs ou d'ALUs. Ainsi, avec des milliers de coeurs au total, ils sont conçus pour effectuer des tâches de façon hautement parallèle. La répartition des transistors est très différente de celle d'un CPU. Un grand nombre de transistors est réservé aux ALUs et très peu à la mémoire cache (mémoire rapide pour les variables locales) et structures de contrôle, voir Fig. 1.8. Cette conception permet au GPU d'être plus rapide en termes de

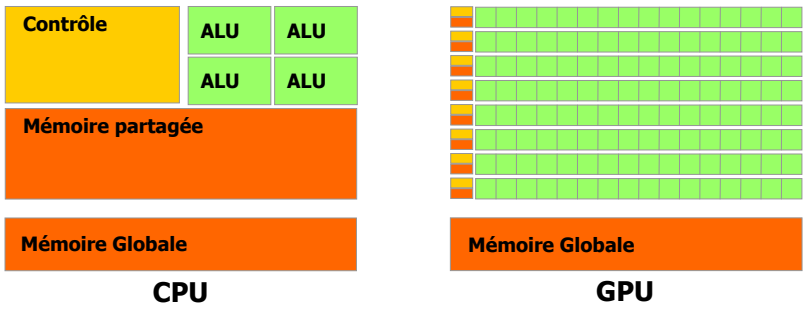


FIGURE 1.8 – Partage de transistors d'un CPU et GPU typiques. Adapté de la figure 3 du Guide de programmation CUDA `https://docs.nvidia.com/cuda/cuda-c-programming-guide/`.

vitesse de calcul brute pour les problèmes mathématiques appropriés. Elle est mesurées en milliards d'opérations en virgule flottante par seconde (GFLOP/s). En l'an 2015, cette métrique a atteint 10200 GFLOP/s en précision de 32 bits pour les GPU les plus performants de NVIDIA, 7.8× plus rapide par rapport aux meilleurs CPU d'Intel® (Santa Clara, Califor-

nie) [95]. De façon correspondante, elle est à 5300 GFLOP/s en précision de 64 bits, 7.6× plus rapide que les CPU d'Intel. Une métrique connexe est la largeur de la bande passante théorique, c'est-à-dire le taux de transfert de données de la mémoire vive vers les ALUs dans des conditions optimales, qui se mesure en giga-octets par seconde (GiB/s). Elle atteint 725 GiB/s pour les meilleurs GPUs de NVIDIA, versus 80 GiB/s pour les meilleurs CPUs d'Intel, soit un avantage de 9×. Ces métriques expliquent en grande partie le succès du GPU en CIP. Cependant, il est important de remarquer que l'utilisation du GPU est en pratique limitée à des problèmes qui se prêtent bien à la parallélisation, et nécessite une programmation experte pour atteindre les facteurs d'accélération théoriques annoncés ici-haut. En conséquence, on se penchera sur quelques concepts importants de programmation pour GPU, ainsi que leur impact sur l'implantation de la reconstruction itérative en TDM.

1.5.4 Parallélisation de problèmes numériques

Quelques questions primordiales s'imposent avant le choix de la programmation parallèle comme solution à un problème numérique [96] :

— le problème à résoudre possède-t-il des tâches exécutables de façon concurrente ? De façon équivalente, le problème peut-il se diviser en sous-problèmes indépendants [à résoudre de façon concurrente] ?

— Si oui, quel pourcentage du temps machine est consacré à ces tâches ? Est-il significatif ?

Lorsque l'on constate que le problème se subdivise bien et une grande partie du temps est consacrée à ces calculs concurrents, la parallélisation est un choix intéressant pour réduire le temps d'exécution. Formellement, l'accélération théorique via parallélisation se calcule via la loi d'Amdahl [97]. Soit une tâche globale dont une fraction du temps d'exécution $p \in [0, 1]$ peut bénéficier d'un facteur d'accélération $s > 1$. Le facteur d'accélération global de l'exécution sera :

$$S(s) = \frac{1}{(1-p) + \frac{p}{s}}. \tag{1.16}$$

Il en résulte qu'à $s = \text{const}$, le plus grand facteur global est obtenu pour un problème où $p = 1$. Pour $p < 1$, l'accélération est limitée par le temps d'exécution non-parallélisable $1 - p$. Il faut aussi réaliser que l'accélération via parallélisation possède un coût objectif : l'achat de l'équipement de calcul approprié et l'énergie dépensée sur son fonctionnement et entretien. Un coût plus subtil est celui du développement de code parallèle et de sa maintenance à long terme.

Une propriété plus spécifique qui influence la programmation parallèle en CIP est le ratio entre le nombre de calculs arithmétiques et le nombre d'accès aux données. Les problèmes sont ainsi classés comme étant *intenses en arithmétique* (*arithmetic-intensive computing*) ou *intenses en données* (*data-intensive computing*). Ce ratio guide le choix des propriétés du matériel selon les caractéristiques recherchées. Pour un problème intense en arithmétique, la

vitesse de calcul brute est le facteur limitant pour le temps d'exécution. Pour un problème intense en données, la bande passante constitue le facteur limitant. Comme mentionné plus haut, le GPU est avantagé du côté des deux métriques. Toutefois, il est important de noter que le GPU est généralement doté de très peu de mémoire cache, ce qui fait en sorte que certains problèmes intenses en données ne sont pas en mesure d'atteindre la bande passante théorique. Ceci résulte en des accélérations moins spectaculaires que pour les problèmes intenses en arithmétique. Cette problématique sera soulevée en détail au chapitre 2.

1.5.5 Plateformes logicielles pour le GPU

La programmation sur GPU peut se faire via diverses interfaces de programmation ou librairies. Les librairies de calcul utilisant le GPU sont conçues pour effectuer des calculs généraux sur de grands ensembles de données : algèbre linéaire (calcul vectoriel et matriciel), transformées mathématiques telles que la transformée de Fourier discrète, analyse statistique, etc. Un exemple bien connu est la Parallel Computing Toolbox™ de Matlab® [98]. Une autre option est le langage OpenCL [93]. Il s'agit d'un langage de haut niveau basé sur la syntaxe du langage de programmation C99, sous licence de logiciel libre. Il possède l'avantage de faire abstraction du matériel et de permettre au développeur d'écrire un code unique qui fonctionne autant sur un CPU multi-coeur que sur des GPU de différents fabricants. L'option privilégiée pour cette thèse est d'utiliser la plateforme CUDA™ de NVIDIA [92]. Cette plateforme comprend des pilotes de matériel graphique permettant l'exécution de calculs généraux sur GPU et une interface de programmation qui est une extension du langage C/C++. L'architecture CUDA est de la classe *instruction unique, données multiples*, connue comme SIMD, de l'anglais *single instruction, multiple data*. Les particularités de cette architecture seront brièvement énoncées ci-après et discutées en détail dans les chapitres 2 et 4.

OpenCL et CUDA étant deux choix intéressants, le choix de CUDA mérite d'être justifié. Étant un langage de haut niveau, OpenCL promet un développement plus rapide du logiciel. De plus, il n'est pas dépendant d'un matériel en particulier, offrant ainsi plus de flexibilité à l'usager final quant au choix du fabricant d'équipement. Toutefois, lorsque l'on cherche de façon prioritaire à maximiser l'accélération par GPU, CUDA est reconnu pour offrir une meilleure performance [99]. Ceci est dû à une liste étoffée d'optimisations de bas niveau intimement liées à l'architecture matérielle de la plateforme NVIDIA. En revanche, le développeur est contraint de suivre un apprentissage extensif et de maintenir ses connaissances à jour. Cette difficulté est partiellement atténuée par la présence de nombreuses librairies CUDA, telle que `Thrust` [100] pour les opérations vectorielles de base, `cuFFT` pour les transformées de Fourier, etc., qui font partie de la trousse de développement logiciel de CUDA. Étant donné que les applications cliniques potentielles sont surtout orientées vers la performance, CUDA a été sélectionné pour développer le logiciel de reconstruction. De plus, comme cette plateforme possède une grande communauté dynamique de développeurs et usagers, le code créé lors de ces travaux promet

d'avoir une durée de vie intéressante.

1.5.6 Programmation CUDA

La caractéristique fondamentale de la plateforme CUDA est l'architecture SIMD. Le développeur écrit donc un ensemble d'instructions à exécuter en parallèle dénommé *kernel*. Ce dernier s'exécute sur un grand ensemble de données selon le même code. Ceci est possible via les *processus légers* (PLs) ou *threads* en anglais. Chaque PL possède un identifiant unique, qui est une variable locale au PL et lui permet d'accéder aux données qu'il doit traiter. Son identifiant lui permet aussi d'exécuter des branchements conditionnels. Les PLs sont groupés en *blocs* s'exécutant sur un même multiprocesseur du GPU. Un ensemble de blocs constitue une grille de tous les PLs.

Le GPU qui supporte CUDA possède différents types de mémoire, illustrés à la Fig. 1.9. Le registre est de très petite taille et contient les variables locales des PLs. La mémoire partagée est visible de tous les PLs du même bloc et permet l'échange rapide d'informations entre PLs. Ces deux types de mémoire sont situés sur la puce du GPU. Ils possèdent une faible *latence* (nombres de cycles de l'horloge pour lire ou écrire une valeur), mais la taille disponible est dans les dizaines de kilooctets par multiprocesseur. Il existe aussi un grand bloc de mémoire externe à la puce, semblable à la mémoire vive d'un ordinateur standard. Sa taille varie typiquement entre un et douze gigaoctets selon le modèle de carte et elle est visible de tous les PLs. Étant externe, la latence d'accès est très élevée. Cette mémoire est subdivisée en mémoire dite globale, pour le stockage général de données, mémoire constante, pour le stockage de données invariables, et mémoire de texture, qui implémente l'interpolation bilinéaire de façon matérielle. La mémoire externe est liée à la puce via une mémoire cache. Cette dernière ne peut pas être contrôlée par l'interface de programmation, mais est plutôt contrôlée par les pilotes du GPU. Enfin, la mémoire de texture est de type spécial : elle est physiquement structurée en 2D et ainsi permet de faire de l'interpolation rapide au moment d'y accéder.

La programmation sur GPU présente plusieurs défis. Une considération essentielle est le risque de situation de compétition ou *race conditions* en anglais. Le développeur doit normalement s'assurer que deux ou plusieurs PLs ne peuvent à aucun moment tenter de mettre à jour une même variable en séquence trop rapprochée et ainsi corrompre sa valeur. Il faut plutôt que chaque variable soit réservée au PL qui en fait usage. Ceci peut se faire via les opérations dites *atomiques* ou bloquantes (qui assurent le blocage de la variable) ou l'utilisation d'une indexation astucieuse des variables qui empêche les accès concurrents. Une autre approche est de quantifier l'impact des mises à jour concurrentes sur la qualité du résultat des calculs, afin de démontrer expérimentalement que la concurrence cause un biais négligeable [101].

D'autres considérations importantes sont liées à la performance. Le développeur doit s'assurer de minimiser le nombre de branchements conditionnels qui divergent dans un même bloc de

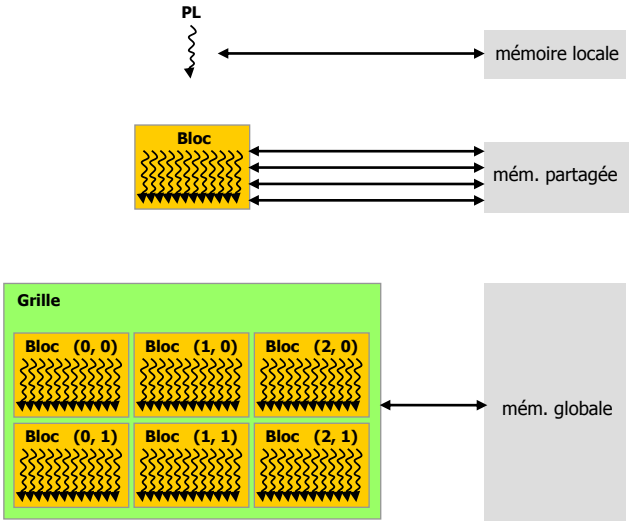


FIGURE 1.9 – Diagramme simplifié des processus légers et types de mémoire chez les GPU compatibles avec CUDA. Adapté de la figure 7 du Guide de programmation CUDA [92].

PLs, car ils sont sérialisés lors de l'exécution. En d'autres termes, le bloc d'instructions assujetti au branchement sera exécuté en série pour chaque PL et seulement une fraction de la puissance de calcul du GPU sera utilisée durant ce temps. Il faut aussi veiller à minimiser le nombre de PLs qui ne font aucun calcul en raison d'une terminaison rapide de leur portion de calcul. Il est aussi important de minimiser les transferts de données superflus ou de les intercaler avec des calculs indépendants. Enfin, il faut porter une attention particulière à l'indexation des variables dans la mémoire globale. La vitesse d'accès de PLs à cette dernière est intimement liée au bon alignement des index de PLs avec les index en mémoire. Il s'agit du concept de *coalescence* de la mémoire, qui sera discuté en détail dans les chapitres 2 et 4.

1.5.7 Parallélisation de la reconstruction itérative

Avec les concepts de base en programmation parallèle et GPU, il est possible d'analyser la stratégie générale de la parallélisation de la reconstruction itérative en TDM. La présence d'une séquence d'itérations impose une limite naturelle au degré de parallélisation d'un algorithme itératif. En pratique, ceci n'empêche pas le développement d'une panoplie d'implantations parallèles [24, 47], car chaque itération est riche en opérations parallélisables. Ainsi, le processus de projection (obtention de projections simulées) se prête bien à la parallélisation, car chaque trajectoire d'atténuation est entièrement indépendante des autres. Pour la même raison, la création de la matrice-système (le calcul des poids w_{ij}) est facilement parallélisable. La rétroprojection constitue une étape plus délicate, car chaque voxel du volume est dépendant des facteurs de correction issus d'une multitude de rayons. Ainsi, la rétroprojection doit avoir recours à des étapes de synchronisation dans le but d'appliquer les facteurs de correction sans

contrevenir à la logique du problème numérique original.

Les opérations de projection et rétroprojection en trois dimensions font de la reconstruction itérative un problème intense en données. Ceci est lié à la taille de la matrice-système et aux mises à jour répétitives des projections estimées et de l'image 3D estimée. Toutefois, comme mentionné ci-haut, les implémentations GPU peuvent quand même produire des facteurs d'accélération intéressants à une fraction du coût des grappes de calcul traditionnelles.

Le choix de la précision numérique pour le stockage de données et l'arithmétique a un impact sur le temps de calcul, mais aussi sur l'accumulation d'erreurs numériques. Généralement, le choix de la précision numérique doit être justifié et validé en fonction des exigences du domaine d'application. En TDM, on a démontré pour quelques algorithmes représentatifs que la précision aussi basse que 16 bits (virgule flottante) pour la représentation des données de projection et du volume à reconstruire a causé des biais négligeables [102], et on s'attend donc que des précisions plus élevées soient des choix valides également.

1.6 Avancées récentes en reconstruction itérative

Avec un bon bagage de solutions mathématiques pour le problème de la reconstruction en TDM accumulé depuis l'introduction de cette modalité, ainsi que des outils de CIP en constante évolution, dont l'introduction des plateformes GPGPU comme CUDA (en 2007) et OpenCL (en 2009), le développement et l'accélération de la RI en TDM a gagné un intérêt sans précédent et continue d'être d'actualité, avec les efforts récents et futurs orientés vers le passage de la reconstruction rapide à la reconstruction en temps réel. Il est donc important de faire un survol des avancées récentes en RI tant du point de vue algorithmique que du point de vue de l'accélération matérielle.

1.6.1 Reconstruction et régularisation en TDM à rayons X

En termes de reconstruction, certaines publications récentes optent pour une statistique de Poisson pour l'atténuation des photons [59, 103–105]. Cette formulation est adéquate pour des fluences de photons plus faibles et constitue le modèle utilisé dans cette recherche. L'intégration des phénomènes physiques complexes en TDM à rayons X sous-tend plusieurs domaines de recherche en soi, dont les plus marquants sont le durcissement de faisceau et la radiation diffusée. L'atténuation des effets de la polychromaticité peut s'effectuer via la décomposition d'Alvarez-Macovski [106, 107] ou des méthodes de nature plus empirique [108, 109]. Un sujet connexe est la réduction d'artéfacts métalliques causée par un durcissement de faisceau extrême et possiblement une perte de signal au détecteur [110–113]. La correction des effets de la radiation diffusée est un autre sujet d'étude important [114, 115]. Une bonne revue générale de la modélisation avancée avec ses différents aspects est faite par Nuyts *et al.* [7]. Ce large spectre de connaissances dépasse le cadre de cette recherche, qui se concentre sur le problème

de base de la TDM à faible dose, ainsi que sur certains aspects avancés de la tomographie optique et de l'optimisation numérique.

Outre le modèle de Poisson, beaucoup de publications utilisent la norme euclidienne de la différence entre les lectures expérimentales et la projection à travers l'image 3D estimée, pour poser le problème de reconstruction et formuler le terme de correspondance aux données de la fonction-objectif [18,21,56,57], supposant donc que μ obéit à la loi normale. Ceci constitue une approximation pour l'imagerie basse-dose, où les lectures de détecteur sont basées sur un faible nombre de photons. Pour ce modèle simplifié, la qualité de l'image résultante repose davantage sur la méthode d'optimisation et le terme de pénalisation. Le terme de pénalisation assure la convergence vers une image visuellement attrayante et généralement plus fidèle à l'objet imagé, tandis que la méthode d'optimisation influence la vitesse de convergence vers cette solution.

Un moment important dans le développement d'algorithmes de reconstruction a été marqué vers l'an 2006 avec l'introduction et l'adaptation subséquente de la régularisation via la minimisation de la variation totale (TV), qui sera discutée en détail à la section 2.4.2. Il s'agit d'un critère de pénalisation basé uniquement sur le voisinage de l'image 3D à filtrer, ce qui lui confère une bonne performance numérique. Le concept à la source est l'échantillonnage compressé, ou *compressed sensing* en anglais, qui relève du traitement de signal et de la restauration d'images [116]. Cette méthode a été adaptée à la TDM [18, 56, 57, 105], où l'on minimise la norme du gradient de l'image 3D estimée, ce qui a pour effet d'éliminer les bruits de haute fréquence et d'épargner les principaux contours de l'image. Cette pénalisation, sans être parfaite, en raison de l'apparition dans les images de zones de densité constante à caractère artificiel, est maintenant devenue un étalon pour la validation de nouvelles approches de régularisation. En fait, de multiples variantes de la méthode adaptées à diverses contraintes de la TDM ont fait leur apparition. Le chapitre 2 de la présente thèse s'inscrit dans cette vague d'adaptations, en offrant une implantation GPU de l'algorithme convexe à sous-ensembles ordonnés avec régularisation TV (OSC-TV). Parmi les autres méthodes dérivées, on retrouve la méthode ART simultanée (SART) avec pénalisation TV [117,118], ainsi que l'optimisation via les fonctions de substitution [119] et la méthode du gradient de Barzilai–Borwein [120]. La méthode TV elle-même a connu des modifications, par exemple la formulation à préservation de contours [121], à pondération adaptative [122], anisotropique [123], généralisée [124,125], aux termes d'ordre supérieur [126], à gradient de la TV [127]. Le remplacement de la norme du gradient par une quasi-norme d'ordre $p, 0 < p < 1$ a également été étudié [128,129]. Plus récemment, la variation totale non-localisée a été utilisée en TDM [130].

Une approche de régularisation concurrente développée durant la même période est la méthode de régularisation par sous-images [131,132], mieux connue comme la méthode NLM, de l'anglais *non-local means*. Son principal avantage par rapport à la méthode TV est que lors de tests sur des fantômes connus, l'erreur entre l'image de référence et l'image reconstruite prend

l'apparence d'un bruit blanc, alors que l'erreur de la méthode TV est structurée. Elle a été étudiée en CBCT basse-dose [133] et en CBCT 4D [134].

Quelques études récentes ont été consacrées aux approches d'optimisation de la fonction-objectif, dans le but d'augmenter le taux de convergence. Un exemple qui réutilise avec élégance plusieurs travaux récents est un algorithme par Kim *et al.* [135]. Il s'agit une méthode basée sur les fonctions de substitution quadratiques séparables (SQS) de *separable quadratic surrogates* en combinaison avec la division du sinogramme en sous-ensembles de projections et accélération via la méthode des moments de Nesterov. Elle a montré une convergence améliorée sous les hypothèses d'une statistique normale des photons sur des données synthétiques et cliniques. Une méthode dérivée implantée sur GPU a été proposée par McGaffin et Fessler [60].

1.6.2 Opérateurs de projection et rétroprojection

Les contributions récentes dans l'étude des opérateurs de projection et rétroprojection sont surtout concentrées sur l'évaluation comparative des différents modèles, ainsi que sur l'accélération matérielle ; néanmoins, il faut souligner quelques approches nouvelles. Une formulation exacte de l'intersection entre une grille de voxels de type cartésien et des faisceaux de forme pyramidale, qui simule la taille finie des pixels, a été proposée par Yao et Leszczynski [136], permettant d'accélérer la convergence de l'algorithme SART et de réduire les artéfacts de haute fréquence. Brokish *et al.* [137,138] ont proposé les opérateurs dits *hiérarchiques*. Il s'agit d'une approche de type *diviser pour régner*, c'est-à-dire de reconstruire l'image par sous-volumes. Chaque sous-volume requiert moins de données de projection pour atteindre un rapport signal-sur-bruit satisfaisant. Implantée sur GPU, elle a permis d'atteindre des facteurs d'accélération jusqu'à 10× par rapport au code GPU non-hiérarchique. Plus récemment, la compression de la matrice-système via l'utilisation des coordonnées cylindriques a été étudiée [67–69]. Elle permet de réduire la matrice-système aux données correspondant à une dizaine de projections, les autres étant simples à recalculer via une symétrie par rotation.

À part ces approches nouvelles, plusieurs optimisations de méthodes bien établies ont été étudiées récemment. En ce qui a trait aux méthodes approximatives, la méthode guidée par la distance [77] a connu beaucoup de succès comme formalisme pour les implantations GPU : par Miao *et al.* [139], Mitra *et al.* [140] et Schlifske *et al.* [141]. L'interpolation de Joseph pour la projection directe a été accélérée via une *formulation généralisée* [142] implantée sur GPU. Quelques études récentes ont comparé différents modèles approximatifs en termes de convergence et de qualité d'image résultante, concluant que les modèles approximatifs avancés étaient préférables à l'interpolation bilinéaire [143], quoique le protocole d'acquisition influençait la justesse relative des modèles [144].

Pour le modèle aux rayons fins, des méthodes accélérées sur GPU ont été récemment proposées : Chou *et al.* [145] ont étudié l'accélération de la projection directe sur GPU via l'utilisation

de la mémoire partagée, constante et de texture. Gao [146] a développé un formalisme à complexité théorique $\mathcal{O}(1)$, c'est-à-dire de temps constant par longueur l_{ij}, pour le calcul de ces coefficients, sans toutefois s'attarder aux opérateurs de projection et rétroprojection en tant que tels. Également, Nguyen et Lee [147] ont proposé une implantation optimisée de la paire projection et rétroprojection via rayons fins en utilisant une approche modifiée de Siddon pour la projection directe et un modèle de *sphéroïde enveloppant* pour le voxel, lors de la rétroprojection.

1.6.3 Reconstruction itérative en tomographie optique

Les premières tentatives d'introduire la RI en tomographie optique pour la dosimétrie 3D sont très récentes. Quelques études ont exploré la reconstruction itérative sur des modèles numériques de tomographes optiques en 2D. L'approche ART a été évaluée pour la tomographie optique par Rankine et Oldham dans le contexte de l'évaluation d'un nouveau protocole d'imagerie avec égalisation approximative de l'indice de réfraction [148]. Les méthodes ART et SART ont été évaluées par Doran et Yattigammana [37] pour la tomographie optique sans fluide d'égalisation de l'indice de réfraction. Dans l'article présenté au chapitre 3, la reconstruction itérative OSC-TV a été évaluée en tomographie optique en utilisant une grille de reconstruction 3D et des données expérimentales, ce qui constitue une contribution originale d'intérêt pour ce domaine. Par la suite, d'autres évaluations de ce genre ont paru, notamment de la technique itérative simultanée (SIRT) [149] de l'anglais *simultaneous iterative reconstruction technique* par Du *et al.* [150] et SART avec régularisation TV [118].

1.7 Description du projet de recherche

Les paragraphes précédents avaient pour but de familiariser le lecteur avec l'utilité, les concepts théoriques et diverses contraintes de la reconstruction en TDM et tomographie optique, ainsi que d'offrir un survol des travaux récents dans ce domaine. Dans le contexte d'une telle effervescence en reconstruction itérative, entre autres sur GPU, il est très important de situer précisément la contribution scientifique propre et l'impact de ce travail de recherche. En fait, il répond à des attentes bien spécifiques et par sa nature constitue une plateforme pour le développement de méthodes de reconstruction plus avancées en TDM.

1.7.1 Problématique

Les examens TDM ou CBCT en imagerie médicale présentent en général un bénéfice net pour le patient [1], car la procédure permet de révéler ou d'exclure des pathologies ou d'améliorer l'alignement d'un patient qui doit subir une thérapie guidée par l'image. Par contre, le risque associé à l'exposition du patient au rayonnement ionisant est plutôt de nature statistique et est toujours à l'étude [1–3]. Du point de vue quantitatif, il est possible d'augmenter le bénéfice via une amélioration de la qualité d'image, par exemple en détectant des lésions de

plus petites tailles ou de plus faible contraste. Le risque potentiel pourrait être réduit via l'obtention d'images de qualité équivalente à dose plus faible. L'étape de la reconstruction a une influence directe sur la qualité d'image 3D, ce qui influence le rapport bénéfice/risque. Dans le cadre de ce projet, on s'intéresse à construire un nouvel algorithme robuste face à une réduction de dose et/ou démontrant une amélioration de la qualité d'image à dose égale, dans le contexte de l'imagerie CBCT. De plus, on procède à une analyse comparative qui met en évidence ses caractéristiques face à des algorithmes semblables et on apporte une attention particulière à la vitesse d'exécution, qui est une contrainte clinique non-négligeable.

En ce qui à trait à la tomographie optique à faisceau conique, cette modalité de lecture de dosimètres 3D à gel est prometteuse en radiothérapie. Elle consiste à imager un échantillon dosimétrique 3D et d'en faire une carte quantitative de la dose. Il s'agit donc d'un transfert d'information qui dégrade la résolution spatiale, la justesse et la précision des mesures brutes. Il est attendu que la reconstruction itérative amène une amélioration significative de la qualité d'image, de la même façon qu'en TDM à rayons X. Toutefois, il y a très peu de littérature consacrée à étudier l'apport de la RI dans ce domaine, et les quelques études existantes sont axées sur des données synthétiques [37, 148], au moment d'entamer le volet optique de cette recherche. Ainsi, une validation sur des données réelles de l'apport de la RI en tomographie optique constitue une tâche importante dans l'amélioration de dosimétrie radiochromique 3D et ultimement la qualité de la radiothérapie externe.

Avec le nouvel algorithme de reconstruction et son évaluation en CBCT et tomographie optique, on a établi une plateforme intéressante pour le développement de nouveaux algorithmes de reconstruction plus sophistiqués, dont la TDM dynamique, ou TDM-4D. Toutefois, l'algorithme original avec matrice-système stockée était limité par la taille de la mémoire vive du GPU et une solution plus flexible était devenue nécessaire. Ainsi, on s'est attardé au problème de génération et de stockage de la matrice-système, un sujet inséparable de la reconstruction en TDM et continuellement étudié dans la littérature. En fait, la modélisation géométrique influence autant les propriétés de convergence des algorithmes itératifs que leur complexité numérique. En général, une représentation plus juste est plus complexe à calculer, et il existe différents compromis entre justesse et vitesse. Un nombre significatif de publications est consacré à l'élaboration de représentations géométriques nouvelles [71, 73, 77, 78, 83], tandis que d'autres étudient l'accélération algorithmique et matérielle des modèles existants [72, 85, 86, 146, 147]. Le troisième volet de cette thèse s'inscrit dans cette dernière catégorie. Comme la RI suppose plusieurs itérations, l'idée de stocker la matrice-système afin d'éviter de recalculer ses éléments à répétition est intuitive. Quelques études y sont consacrées [87, 88]. Dans la réalité, la bande passante du matériel peut limiter la performance d'une telle approche, rendant le re-calcul des coefficients à la volée plus avantageux. Ainsi, l'ajout d'une comparaison rigoureuse entre le stockage de la matrice-système et le traçage à la volée apparaît important pour le domaine de la RI. Comme l'avantage du GPU sur le CPU dans la reconstruction en TDM est bien

établi [24, 47], l'étude présentée ici est faite uniquement sur GPU. La motivation est donc de bien démontrer les avantages et les limites de ces techniques.

1.7.2 Objectifs de la thèse

L'élaboration d'un algorithme de reconstruction itératif rapide en tomodensitométrie à faisceau conique et l'évaluation de ses propriétés constitue le principal objectif du projet de doctorat. Le projet de recherche se subdivise comme suit :

1. Élaboration d'un algorithme de reconstruction itératif implanté sur GPU et son évaluation en tomodensitométrie à faisceau conique embarquée en radiothérapie externe.
2. Évaluation de cet algorithme en tomographie optique à faisceau conique pour les gels radiochromiques.
3. Implantation et évaluation comparative de diverses méthodes de calcul et de stockage de la matrice-système pour la reconstruction itérative.

La première étape consiste à élaborer un algorithme de reconstruction. Cette approche nouvelle est construite en combinant une fonction-objectif basée sur un modèle physique, avec un bon équilibre entre justesse et complexité numérique, un terme de régularisation avec préservation des contours, une approche d'optimisation de la fonction-objectif et une représentation géométrique. L'algorithme est d'abord évalué sur des données de projection simulées, afin de mettre en évidence les propriétés de convergence de l'algorithme. Une analyse comparative avec quelques algorithmes bien connus dans le domaine s'impose pour démontrer sa pertinence pour la recherche future et la clinique. Ensuite, il est testé sur des données réelles de CBCT, dans le but d'obtenir un bref aperçu de son potentiel clinique. L'optimisation du logiciel pour une exécution rapide sur GPU est un autre volet intégral de cette étape.

La seconde étape consiste à étendre le champ d'applications de l'algorithme vers la tomographie optique à faisceau conique. Cette étude comble un certain vide dans la littérature scientifique en ce qui concerne la RI en tomographie optique et met évidence une amélioration considérable de la qualité d'image. De plus, cette étude fait ressortir les défis de l'imagerie quantitative dans le domaine, dont les effets de la polychromaticité sur la reconstruction 3D.

La troisième étape consiste à développer, optimiser et évaluer les approches de calcul et de stockage de la matrice-système sur GPU. On explore les avantages et les contraintes du traçage à la volée et du stockage complet ou partiel de cette matrice dont la taille, sans compression, dépasse la capacité de la mémoire vive des ordinateurs typiques. Le stockage de la matrice-système requiert également une étude des symétries géométriques susceptibles d'augmenter son taux de compression. On évalue la vitesse d'exécution des opérateurs de projection et rétroprojection avec les diverses stratégies, ce qui constitue un apport scientifique intéressant pour la RI en général. De plus, on évalue la performance particulièrement pour l'algorithme de

reconstruction développé dans le cadre de la recherche. Comme l'implantation GPU développée dans le cadre de cette recherche constitue une plateforme pour l'élaboration d'algorithmes plus complexes et spécialisés, la flexibilité du logiciel face à des contraintes d'application est également discutée.

Ces trois grands objectifs sont traités dans les chapitres 2, 3 et 4 respectivement. Il s'agit de deux articles scientifiques publiés et d'un article soumis. Pour compléter la discussion des effets de la polychromaticité sur la quantification en tomographie optique, l'annexe A présente l'effet de l'aplatissement du spectre lumineux. Dans le but de compléter l'analyse de la qualité d'image, l'annexe B discute de l'effet de la régularisation sur les structures de bas contraste. Enfin, en complément à la méthode de la matrice-système compressée, le concept des symétries géométriques en TDM à faisceau conique est décrit en détail dans l'annexe C.

Chapitre 2

GPU-Accelerated Regularized Iterative Reconstruction for Few-view Cone Beam CT

Dmitri Matenine[1], Yves Goussard[2] et Philippe Després[1,3,4]

[1] *Département de physique, de génie physique et d'optique, Université Laval, Québec (Québec), Canada*

[2] *Département de génie électrique / Institut de génie biomédical, École Polytechnique de Montréal, Montréal (Québec), Canada*

[3] *Centre de recherche sur le cancer, Université Laval, Québec (Québec), Canada*

[4] *Département de radio-oncologie et Centre de recherche du CHU de Québec (Québec) Canada.*

2.1 Résumé

Cet article propose une méthode de reconstruction itérative conçue pour la TDM à faisceau conique basée sur un modèle physique de l'atténuation des rayons X. Cette approche a le but de produire des images de TDM fidèles avec peu de projections, ainsi que des temps de calcul cliniquement acceptables. Cette méthode combine la reconstruction de type convexe à sous-ensembles ordonnés (OSC) et la régularisation via la minimisation de la variation totale (TV) et est désignée OSC-TV. Le nombre de sous-ensembles est progressivement réduit au fil des itérations dans le but d'assurer la performance optimale de la régularisation. Compte tenu de la complexité numérique de cet algorithme, il a été implanté sur matériel graphique, une plate-forme de calcul hautement parallèle. Des reconstructions de données simulées, ainsi que d'un sinogramme de TDM à faisceau conique d'un pelvis humain ont été effectuées afin d'analyser la qualité d'image. En termes de convergence, OSC-TV offre une bonne qualité

d'image pour les acquisitions à peu de projections et surpasse la méthode de projections sur les ensembles convexes avec régularisation TV (POCS-TV). Cette méthode constitue aussi une alternative intéressante à la rétroprojection filtrée appliquée à un sinogramme complet. Les temps de reconstruction varient entre une et deux minutes et sont compatibles avec le rythme du travail clinique pour les applications qui ne requirent pas la visualisation en temps réel. Compte tenu de sa capacité de réduction de dose, cet algorithme pourrait être très utile en clinique, surtout pour réduire les risques radiologiques chez les patients qui sont soumis à des examens multiples.

2.2 Abstract

Purpose : The present work proposes an iterative reconstruction technique designed for X-ray transmission computed tomography (CT). The main objective is to provide a model-based solution to the cone-beam CT reconstruction problem, yielding accurate low-dose images via few-views acquisitions in clinically acceptable time frames.

Methods : The proposed technique combines a modified ordered subsets convex (OSC) algorithm and the total variation minimization (TV) regularization technique and is called OSC-TV. The number of subsets of each OSC iteration follows a reduction pattern in order to ensure the best performance of the regularization method. Considering the high computational cost of the algorithm, it is implemented on a graphics processing unit, using parallelization to accelerate computations.

Results : The reconstructions were performed on computer-simulated as well as human pelvic cone-beam CT projection data and image quality was assessed. In terms of convergence and image quality, OSC-TV performs well in reconstruction of low-dose cone-beam CT data obtained via a few-view acquisition protocol. It compares favorably to the few-view TV-regularized projections onto convex sets (POCS-TV) algorithm. It also appears to be a viable alternative to full-dataset filtered backprojection. Execution times are of one to two minutes and are compatible with the typical clinical workflow for non-real-time applications.

Conclusions : Considering the image quality and execution times, this method may be useful for reconstruction of low-dose clinical acquisitions. It may be of particular benefit to patients who undergo multiple acquisitions by reducing the overall imaging radiation dose and associated risks.

keywords : cone-beam CT, few-view CT, iterative reconstruction, total variation minimization, GPGPU

2.3 Introduction

X-ray CT is a widely used medical volumetric imaging modality. It relies on multiple X-ray projections of the subject to mathematically reconstruct a 3D distribution of the attenuation coefficients within the subject. This modality has a number of specialized clinical applications and is still under active development [18]. Currently, CT radiation accounts for a large proportion of the ionizing radiation dose to the population, with potential risks for the patient, including radiation-induced cancers [3]. Therefore, the reduction of CT radiation dose is an important issue and has recently attracted the attention of the scientific community, the media and different government regulation agencies. Similarly, cone-beam CT (CBCT) is used in a number of imaging applications, e.g., patient positioning and treatment planning in external beam radiation therapy. Each patient may undergo several CBCT acquisitions during treatment, increasing the total radiation dose to healthy tissues, therefore increasing tissue damage and secondary cancer risk [151].

The capacity to reduce dose depends in part on the algorithm used for image reconstruction. Most clinical scanners still use algorithms dating back to the 1980's, when computational resources were very limited by today's standards [18]. One of the most widely used algorithms for cone-beam CT (CBCT) applications, for instance, was published in 1984 [4]. This analytical technique, dubbed FDK, enjoys a relatively low computational cost. However, FDK requires a high number of projection views to yield quality reconstructions. This algorithm does not incorporate a sophisticated denoising scheme, so each projection requires a relatively high X-ray exposure.

Iterative reconstruction techniques offer a realistic model of X-ray photon attenuation in the subject and are potentially superior to analytical approaches for low-dose acquisitions [18]. These algorithms may be classified according to the objective function that is iteratively maximized and the method employed to maximize the objective. The ART [46] family minimizes the l^2 norm of the difference between the real and estimated projection data, under the assumption that the photon counts at the detector follow a normal statistical distribution. Other algorithms [51, 53, 152] assume that the photon counts follow a Poisson distribution. The latter is more accurate than a normal distribution at low photon counts encountered in low-dose imaging. It was retained in the present work to preserve image quality despite the low photon counts of some projection readings. The optimum of the objective function is obtained either via various line searches, gradient descent or expectation-maximization. The latter is used in the ordered subsets convex (OSC) algorithm [54] which is the basis of the method proposed here. Compared to the original convex algorithm [53], the ordered subsets (OS) approach accelerates the computation by one to two orders of magnitude, at the cost of introducing image bias [153]. Considering the clinical requirement for fast reconstruction, the ordered subsets framework was used in this work.

Regularization is necessary to address the ill-posed character of the problem, namely the incomplete dataset and noise. The regularization is also based on the hypotheses concerning noise and *a priori* knowledge of the image properties. Current research is rich in regularization criteria, aimed at offering edge preservation and noise reduction. A relatively robust criterion is the sparseness of the magnitude of the image gradient. This criterion was used here through the total variation (TV) minimization technique [56]. Even though combinations of the expectation-maximization and TV algorithms were proposed [105], the specific case of OSC reconstruction combined to TV regularization was not studied. The present paper modifies the OSC algorithm in an innovative manner to suit the particularities of TV regularization, improving image accuracy for a few-view acquisition.

Iterative reconstruction is usually characterized by a high computational cost, both in arithmetic operations and memory bandwidth [47], even when using the ordered subsets approach. Strategies using cylindrical geometries may alleviate this problem, although convergence properties still require investigation [67]. The sequential and memory-intensive nature of iterative algorithms makes them poorly suited, *a priori*, for parallel implementation on Graphics Processing Units, whereas analytical approaches have benefited enormously from these highly-parallel devices [154]. Nevertheless, the steady improvement of the compute capabilities of GPUs makes them increasingly suitable for iterative reconstruction in CT, e.g., reaching reconstruction times in the order of two minutes for a few-view set of 40 CBCT projections [20]. The method proposed here was implemented on the GPU and the code was carefully optimized according to the algorithm's specifics. It was tested on computer-simulated as well as clinical projection data and compared to the results of FDK and TV-regularized projections onto convex sets (POCS-TV) [56] algorithms.

2.4 Materials and Methods

2.4.1 Photon attenuation model

The attenuation model assumes a mono-energetic photon beam passing through an attenuating medium :

$$Y_i = d_i e^{-t_i}, \tag{2.1}$$

where i denotes a detector reading index, Y_i is the detector photon count for that reading, d_i is the incident photon count, $t_i = \sum_j l_{ij}\mu_j$ is the total attenuation along the ray path. The latter is based on discrete length paths l_{ij} and linear attenuation coefficients μ_j, where j denotes voxel indices. Intrinsic characteristics of a CT-scanner affecting the dose, e.g., the quantum detection efficiency of the detector panel were not taken into account. Also, the detector intensity reading was considered as a sum of photon counts. The hypotheses of a mono-energetic beam and a photon-counting detector are both over-simplifying with regard to

most clinical CT-scanners, which use a poly-energetic beam and energy-integrating detectors. However, algorithmic corrections for polychromaticity and detector properties, which deserve a separate literature review and investigation, are beyond the scope of this paper.

2.4.2 Reconstruction algorithm basics

The goal of the current work is to reduce ionizing radiation dose due to CBCT imaging. Two approaches are considered here : dose reduction mainly through the reduction of the number of projections, and, to a much lesser extent, through the reduction of the number of incident photons per projection. The former renders the problem to solve under-determined, while the latter introduces noise. Regularization is an effective way to constraint an under-determined problem and an appropriate photon attenuation model is capable of reducing the image degradation due to noise. The proposed algorithm iteratively maximizes a two-term criterion :

$$F(\mu) = L(\mu) - \lambda||\mu||_{TV}, \tag{2.2}$$

where $L(\mu)$ is a Poisson log-likelihood of the image estimate μ and serves as the data-correspondence criterion. The regularization term $||\mu||_{TV}$ denotes the total variation of the image estimate. It is affected by a weighting factor λ, which determines the extent of the regularization process, and is not explicitly used in this work, as will be shown in section 2.4.3. This second term is subtracted because the TV-norm must be minimized while the log-likelihood must be maximized. In practice, the two terms are optimized alternately, which yields an approximate solution to Eq. 2.2. The motivation for such an approximation is the following : the parameters of the TV-norm minimization step require the result of the previous data-correspondence step [56]. Also, executed separately, the two quite dissimilar computing tasks can undergo separate optimization for the GPU.

The data-correspondence criterion assumes that detector readings are Poisson-distributed photon counts, and maximizes the following log-likelihood function :

$$L(\mu) = -\sum_i (d_i e^{-t_i} + Y_i t_i). \tag{2.3}$$

This function is known to possess a unique maximum [53] with respect to μ, which corresponds to the best fit of the image estimate to projection data Y. The expression of the maximum is a transcendental function. Therefore, as suggested by Lange and Fessler [53], a single iteration of the Newton method is used at each estimate update, yielding the Convex algorithm :

$$\mu_j^{(n+1)} = \mu_j^{(n)} + \mu_j^{(n)} \frac{\sum_i l_{ij} \left[\overline{y}_i^{(n)} - Y_i\right]}{\sum_i l_{ij} t_i^{(n)} \overline{y}_i^{(n)}}, \tag{2.4}$$

where n denotes the reconstruction iteration number and $\overline{y}_i = d_i e^{-t_i^{(n)}}$ is the estimated photon count. It is possible that $\mu_j^{(n+1)} < 0$ even if $\mu_j^{(n)} > 0$, so the non-negativity of μ should be

enforced to respect the physics of the problem [53]. The convergence of this algorithm is known to be slow, since it is equivalent to a steepest descent method. To accelerate computations, the ordered subsets convex (OSC) algorithm was used in this work. Despite the bias introduced into the solution by the OS formulation, it yields a relatively fast convergence to an acceptable solution and uses a low number of well documented and quite predictable parameters [155]. OSC updates μ after projecting and back-projecting only a few (e.g., two) projections in sequence [54, 153] :

$$\mu_{s+1}^{(n)} = \mu_s^{(n)} + \mu_s^{(n)} \frac{\sum_{i \in S(s)} l_{ij} \left[\overline{y}_i^{(n)} - Y_i \right]}{\sum_{i \in S(s)} l_{ij} t_i^{(n)} \overline{y}_i^{(n)}}, \tag{2.5}$$

where s is the index of the subset and $S(s)$ is the function generating the subsets of rays associated with the selected projections. A full step is completed when all projections have been used to update $\mu^{(n)}$. For every sub-step, Kamphuis and Beekman [54] suggest to select the projection subset with the greatest *angular distance* from all the previously used projections. A slightly simplified subsets ordering approach is implemented in this work. It is performed by picking one subset at a time, in the middle of the largest group of adjacent "not-yet-picked" subsets, until all subsets have been picked.

Each resulting image of a full OSC iteration step is regularized within the image space by the total variation (TV) minimization technique [56]. This approach has the advantage of performing computations in image space only, yielding low computing times and easy parameter tuning, both desirable in a clinical setting. However, its simplistic objective function leads to the appearance of characteristic patch-like artifacts of constant density and to the loss of small structures within the image. For this reason, this particular method has been challenged by more sophisticated recent developments, such as adaptive-steepest-descent TV minimization [57] and higher order TV minimization [156]. The approach selected here may be seen as a special case of advanced TV methods, using fewer free parameters than the latter. The proposed method could be replaced by a more recent development without adversely affecting computation time, as will be shown in section 2.4.4.

The total variation of an image is a scalar number defined as the sum of the components of the norm of the gradient of the image. For a two-dimensional (2D) image :

$$\begin{aligned} ||\mu||_{TV} &\equiv \sum_{a,b} |\vec{\nabla} \mu_{a,b}| \\ &= \sum_{a,b} \sqrt{(\mu_{a,b} - \mu_{a-1,b})^2 + (\mu_{a,b} - \mu_{a,b-1})^2}, \end{aligned} \tag{2.6}$$

where a, b are the pixel indices. The image gradient is not defined at the image boundaries and is usually forced to zero. As suggested by Sidky *et al.* [56], the total variation is minimized by gradient descent. Let $v_{a,b} = \partial ||\mu||_{TV} / \partial \mu_{a,b}$ denote the local value of the TV gradient. Image regularization is performed iteratively by subtracting the matrix v from the image estimate

and calculating the new gradient :

$$\mu_{a,b}^{(q+1)} = \mu_{a,b}^{(q)} - cd_A \frac{v_{a,b}^{(q)}}{\sqrt{\sum_{a,b} v_{a,b}^2}}, \tag{2.7}$$

where q denotes the TV iteration index, c is an empirical constant and the scaling coefficient $d_A \equiv \sqrt{\sum_{a,b} \left(\mu_{a,b}^{(n)} - \mu_{a,b}^{(n-1)}\right)^2}$ is the norm of the difference between the images at iterations n and $n-1$, further referred to as *image displacement*. With this correction factor, the image is regularized proportionally to the changes induced by the latest OSC step. In this work, the TV regularization parameters were fixed empirically. For the synthetic images, based on the lowest difference between the final estimate and the phantom, the optimal values were found to be $c = 0.2$ and $q \in [1, 20]$ after several trials. For clinical data, in order to enhance the spatial resolution of realistic anatomical features, c was rather set to 0.02 for OSC-TV. One can note that image displacement cannot be calculated after only one OSC iteration, since no previous step is available. As a consequence, no regularization is performed on the result of the first OSC iteration in the present work. Starting at the second OSC iteration, and for all the subsequent ones, the image displacement is computed and the regularization is applied. As mentioned before, OSC-TV, just like the original TV approach [56], does not solve the typical penalized log-likelihood problem where the data-correspondence term and the penalty function are integrated in a single objective function. This is an intrinsic property of the algorithm itself, which requires a complete data-correspondence step in order to scale the regularization step using d_A. Thus, the integration of the penalty term into the objective is not possible and regularization is alternated with the OSC step.

2.4.3 Modified OSC-TV reconstruction

It is known that the ordered subsets approach introduces image bias. Also, the step size when applying the TV gradient at each regularization iteration (Eq. 2.7) depends on the current value of d_A. In consequence, the number of subsets used for each OSC iteration must be carefully adjusted from one iteration to the next. Originally, Beekman *et al.* [153] suggested to perform the OSC reconstruction using a large number of subsets $|S|_i$ (e.g., 100 subsets of 2 projections each for a 200-projection acquisition) to achieve maximum computation acceleration. Then the number of subsets is reduced to a value $|S|_f$ for the final iteration only (typically $|S|_f$=$|S|_i/10$), in order to reduce the bias. The original approach, further denoted "step function", results in an abrupt drop of the image displacement only at the end of the reconstruction. To achieve better results when using the TV technique [56], the d_A corresponding to each data-correspondence step must be gradually reduced. To satisfy a requirement such as this, a hybrid general-purpose optimization algorithm was proposed by Bertsekas [157]. A variable parameter of his algorithm, say $\alpha \in \mathbb{R}^+$, controls the likeness of the data correspondence step to an *incremental gradient method* step, or to a *steepest descent method* step, in our case equivalent to OSC with many subsets or to EM (one subset), respectively. Bertsekas clearly

demonstrates the advantage of using the acceleration provided by OS when the estimate is far from the solution, then gradually passing to EM close to the solution. However, the original algorithm is very computationally intensive to accommodate any real value of α. In this paper, a similar effect is achieved by simply reducing the number of subsets after each iteration. The subset numbers are integers, chosen to closely follow some continuous decreasing function. The continuous function was chosen empirically as a decreasing power function, scaled and translated in order to start at the maximum number of subsets and finish at the minimum. The actual number of subsets is computed as follows :

$$|S|^{(n)} = \text{round}\left[\frac{|S|_i - |S|_f}{(n_{\max}-1)^p}\,(n_{\max}-1-n)^p + |S|_f\right], \tag{2.8}$$

with $n \in [0, n_{\max}-1]$ and $p \in (0, 1]$. The power function and the associated p interval result in a non-concave graph, varying from a shape close to the classic step function with $p \ll 1$ to a linear shape for $p = 1$. The motivation for a non-concave decrease of $|S|$ is to take maximal advantage of the OS acceleration. Therefore, it is appropriate to first perform a large number of highly accelerated OSC-TV steps and then a few moderately accelerated OSC-TV steps. Conversely, a concave $|S|$ decrease would use very few fully accelerated steps, then applying minor corrections to an estimate that is still far from the OSC-TV optimum.

The variation of p samples the available paths between the linear and step functions, which may be viewed as reasonable boundaries. The gradual reduction of $|S|$ using the above function is the proposed innovation, which suits well the TV regularization. The flowchart of the OSC-TV algorithm is presented in Fig. 2.1.

It should be duly noted that subsets number reduction was explored by Ahn *et al.* [55] in more general terms of global convergence and convergence speed. In contrast, the present study is aimed precisely at optimizing the OSC and TV combination. The behavior of the new approach proposed here is quite predictable, which is a desirable property in clinic, and the preferred value of p is situated in a narrow interval as discussed in section 2.5.

2.4.4 GPU implementation

The combined OSC-TV algorithm was implemented for execution on NVIDIA® (Santa-Clara, California) GPUs using the CUDA™ programming interface. A CUDA-enabled chip possesses hundreds of arithmetic logic units (ALU's), grouped into independent blocks, to perform computations in parallel. Each block possesses a limited amount of registers and fast shared memory, which are dedicated to the block, and are therefore not usable by other blocks. There is also a large global memory that is accessible by any ALU, but access is significantly slower than for the former types. Each ALU executes a *thread*. Similarly, blocks of ALU's execute *thread blocks*. The code for all threads is the same, but the local variables within each thread are individually allocated. The main limitations of the architecture are the following : the

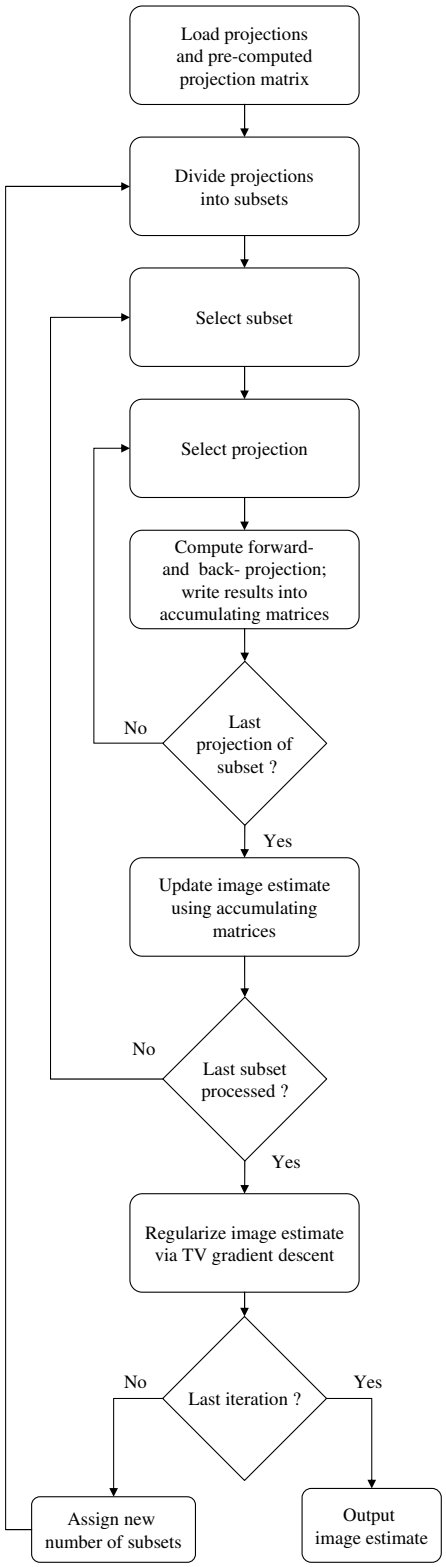


FIGURE 2.1 – Flowchart for the OSC-TV algorithm. For brevity, not shown : the decision step to skip the TV regularization after the first OSC step, as well as the details of the GPU implementation.

control sequence is the same for all threads and divergent conditional statements may significantly reduce performance ; the parallel execution of threads is not automatically monitored for memory access conflicts, the latter being eliminated only through careful programming. So-called atomic operations provide conflict-free memory access, at the cost of a greater latency. Another problem concerns idle threads : a thread may have completed the required computations faster than other threads and stay idle, waiting for other threads to complete, reducing the overall system occupancy. To overcome this problem, the threads should be launched by groups that are expected to have similar computation times. The parallelization strategy is dependent on the problem to solve and is significantly different for the OSC and TV steps of the reconstruction. The OSC step is performed as follows : a particular subset is selected, then a projection part of this subset is fixed. For this projection, a particular "group of rays" of the projection which can be safely processed in parallel is selected (more on this below). These rays are processed in one *CUDA kernel*, i.e., a single parallel computation launch. One kernel integrates the rays' projection and backprojection, however, the results are stored in accumulating matrices until the entire subset has been processed. Then, the image estimate update can take place.

As required by the CUDA architecture, the kernel is executed as a number of blocks, each of the latter taking in charge one or several ray paths. The number of threads per block is equal to the maximum number of voxels crossed by any ray of a projection. To forward-project a ray, the corresponding l_{ij} and $\mu_s^{(n)}$ vectors and the references to the voxels are loaded into the shared memory. The forward projection is computed within the same kernel as the dot product of these two vectors. One block may treat as many rays as the cache memory capacity permits – one to eight depending on the simulation geometry and GPU hardware used.

The backprojection is part of the kernel above, so the previously loaded vectors are reused, reducing execution time. For every image voxel crossed by a particular ray, the elements of accumulating matrices storing the numerator and denominator in the Eq. (2.5) are incremented by the ray's contribution. This step requires read/write operations not to be performed concurrently for the same memory locations. To comply with this requirement, one kernel processes rays from a single projection, to avoid inter-crossing rays that would result in memory conflicts. For a single projection, it is trivial that infinitely thin radiological paths do not intersect physically in the volume. However, rays directed towards two or more contiguous detector bins can travel close enough to each other to cross the same voxel. Using a worst-case scenario, a computation method to find a "safe" distance between detector bins was established. It considers an angular orientation of the volume such that the distance between crossing rays is minimum and the voxel cross-sectional area is maximum. For example, for the geometry used to generate the synthetic data in this work, each 3^{rd} bin on a detector row or column receives rays that cross distinct voxels at any given moment. Thus, for the 2D detector plane, $1/9^{\text{th}}$ of the rays of a projection can be backprojected in parallel, yielding the aforementioned "group

of rays" treated by the kernel.

The main bottleneck within the integrated projection-backprojection kernel is the voxel update, a so-called "scatter" [47] operation which writes to random memory locations. In order to improve the code performance, the matrices that store the numerator and denominator of Eq. (2.5) are organized in *subvolumes*, of size 4×4×2 voxels for example, stored in sequence. The CUDA global memory read and write operations fetch several adjacent bytes per operation, so an entire subvolume is fetched at unitary time cost. The process is illustrated in Fig. 2.2. The simulated rays necessarily cross neighboring voxels connected by at least one face, so several values will be written to the accumulating matrices in unitary time.

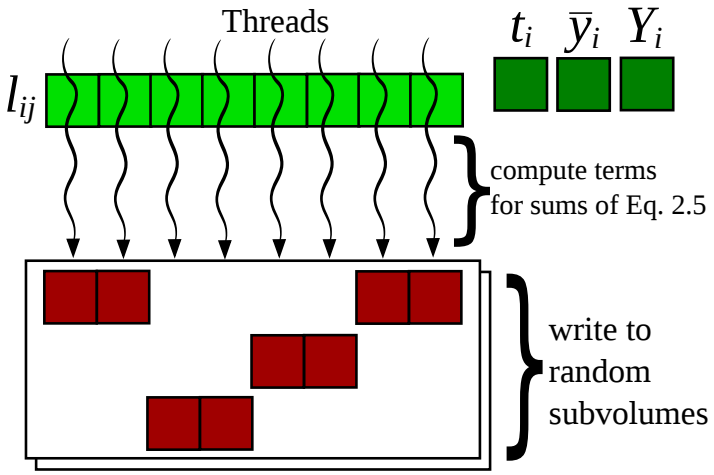


FIGURE 2.2 – Simplified diagram of the parallel execution of the backprojection (M-step). Here, one thread block back-projects along one ray path. The intersection lengths l_{ij} are already loaded into shared memory, as well as the result of the forward projection. The results of the backprojection are written into two accumulating matrices corresponding to the numerator and denominator in Eq. 2.5. The memory coalescence of the *write* operation is limited to two adjacent voxels in this illustration.

A potential bottleneck for GPU computing is the latency of data transfers between the host (PC) and device (GPU) memory, limited by the throughput of the PCI connection. In the present work, repetitive data transfers are avoided altogether. The projection matrix is precomputed and is the largest data structure stored in memory. It is stored in three data vectors, one containing the attenuation lengths for each ray (l_{ij}), another containing the references to the crossed voxels, and one to record individual rays' endpoints, yielding a typical sparse matrix storage scheme. Three types of symmetries of the projection geometry were implemented, imposing some restrictions on the acquisition protocol. First, the rotational symmetry allows one to store only the projections of the first 90°, with the remaining voxel references easily computed "on the fly". The projections are then required to be equally spaced, and their number is required to be a multiple of four, for an angular range of 360°. This allows for a 4×

matrix compression for a 360° range, and similarly, 3× for 270° range and so forth. Another symmetry is the reflection within the x-y (axial) plane, so voxel references from the $[0°, 45°]$ interval are converted to references for the $(45°, 90°]$ interval. This symmetry, in its simplest form, requires the detector panel to be parallel to one face of the reconstructed volume at the first projection and provides another 2× compression. The reflection symmetry relative to the z-axis was also implemented, with the constraint that the central axial plane of the reconstructed volume is aligned with the long axis of the detector array, yielding again 2× compression. The latter two symmetries can still be applied if the detector is slightly misaligned by an integer number of detector pixels. Depending on the acquisition properties, the combined effect is to reduce the projection matrix size up to $4 \times 2 \times 2 = 16$ times, e.g. from 70 GiB to 4.5 GiB of data for the simulated scanner geometry presented in section 2.5. The worst case scenario, for which compression is impossible, is an acquisition that starts at a random angle, is done at highly irregular angular intervals, with the detector misaligned by a non-integer number of detector pixels in both directions. In consequence, the current results in terms of maximum reconstruction grid size are not applicable to an arbitrary scanner geometry. Regarding the practical aspect of the precomputed matrix method, it is assumed that a collection of projection matrices is precomputed and stored on disk to be retrieved on demand. Each matrix is valid for a specific acquisition protocol and reconstruction grid geometry. For example, using a one-terabyte disk, one can store an order of 200 different projection matrices for fast reconstruction. Loading large datasets from disk can require non-negligible time. However, this step can be performed concurrently with image acquisition, therefore it is not included in the total reconstruction times presented in this paper. See section 2.5.6 for details.

The TV regularization was performed in 3D using the following parallelization scheme. The TV gradient is computed for each volume voxel, using thread blocks with a typical size of 16×16×4. The image estimate is similarly subdivided into blocks and loaded into shared memory. Each thread computes the 3D TV gradient $v_{a,b,c}$ after reading the μ values of neighboring voxels stored in shared memory. Therefore, to compute the gradient at the boundaries of a block, a set of additional neighbors is necessary, resulting in cached volumes of 18×18×6 voxels per block. It should also be noted that the boundary voxels of the full image estimate naturally lack some of the neighbors necessary to compute the gradient, in which case it is set to zero.

A more complex objective function and its gradient may be integrated in the current code, as a replacement to the simplistic TV gradient. Added arithmetic operations would not substantially modify the execution time, which is largely influenced by memory operations. Nevertheless, the cardinality of the neighborhood system, i.e., number of neighboring values required to compute the update value for one volume voxel, must not increase significantly. To properly reconstruct thick slices often employed in clinical protocols, the finite differences used to compute the gradient v are normalized by the distances between the centers of the corresponding non-cubic voxels. For both the OSC and TV steps, common operations like matrix sums, sums

of all elements of a matrix or multiplication by a scalar value were performed using the Thrust for CUDA [100] library.

2.5 Results and discussion

2.5.1 Projection data

The present study is based on computer-simulated input and clinical acquisitions. The head section of the NURBS-based XCAT phantom [158] was used to generate human-like volumetric attenuation data. This phantom was selected for a number of reasons. It includes many realistic anatomical features, including small structures, offers a continuous (non-voxelized) X-ray projection software to simulate acquisitions and gives full control over acquisition parameters. The latter is very convenient compared to experimental data, which is affected by many physical phenomena that are not necessarily of interest for this particular study.

The numerical phantom contained the linear attenuation coefficients of the tissues at a 70 keV photon energy. The projections were obtained by raytracing through the continuous phantom using the XCAT `ct_project` projector, with four rays per detector element, originating from a monoenergetic point source set to 70 keV. The same projector was also used to simulate Poisson noise in the projection data, based on the mAs setting. Thanks to the continuous tissue boundaries and raytracing on an oversampled detector grid, this model is different from the one used for creating the projection matrix for the reconstruction. Therefore, a situation of *inverse crime* is unlikely in this work. Here, such a situation would occur if the direct model used for simulating the measurements was identical to the one used for image reconstruction ; this may provide biased indications on how the method performs on actual data. See Kaipio and Somersalo [159] for details.

The simulated projection parameters were similar to a standard-dose head acquisition on a Varian® (Palo Alto, California) OBI 1.4 CBCT system [160]. The tube current was simulated as 0.4 mAs per projection according to the aforementioned protocol, yielding a SNR of 51.4 dB for log-transformed projection data (t_i). This figure is somewhat high and is due to a simulated detector that deterministically absorbs all the incident radiation. The number of views, originally at 360, was reduced to 200, thus simulating a dose reduction of 45% compared to the standard protocol. If implemented in a clinical context, a smaller number of projections acquired at standard dose levels carries two major advantages. First, it reduces the projection matrix size, accelerating computations. Second, the electronic noise introduced by a real detection system, which is significant only for readings obtained at low X ray exposures [161], may be safely ignored at standard dose. Accordingly, the electronic noise is not simulated in this study. The gantry angular range was set to 360° (instead of the OBI's 200° for the head protocol) for all acquisitions in order to achieve similar experimental conditions for all the reconstructions. The use of a reduced angular range would either degrade the FDK image

significantly or require an algorithmic correction.

The clinical acquisition was performed on a Varian OBI 1.4 system. It used a standard pelvis protocol [160], showing the anatomy surrounding a male patient's prostate. The original acquisition contained 688 projections over 360°, with approximately 1.0 mAs per projection. A subset of 172 projections was used to simulate a four-fold dose reduction. Due to a half-fan acquisition geometry, the axial reflection symmetry was unusable, increasing the projection matrix size and therefore limiting the z-span of the reconstruction grid. It should be noted that the acquired projections' angles are not regularly spaced. If compared to a regular angles distribution, the mean absolute deviation is $(0.235 \pm 0.179)^\circ$ for the 172 projections set. In order to maintain a high compression level, this dataset was reconstructed assuming regularly spaced angles. Central slice reconstructions of synthetic data showed that this approximation did not degrade image quality.

2.5.2 Reconstruction grid geometry

For simulated projections, the source-isocenter and the detector-isocenter distances were simulated to be 100 cm and 56 cm respectively. The detector panel was simulated as a 39.7×14.9 cm^2 flat panel. The resolution was set to 512×192 isometric pixels. The reconstructed field of view (FOV) was of 25.6 cm with a total z-span of 11.2 cm. However, the extremities on the z-axis were affected by a truncation artifact, yielding a mostly artifact-free reconstruction on a z-span of 8.4 cm. The grid was of $384 \times 384 \times 192$ cubic voxels with 0.667 mm sides, of which 126 slices were mostly free of the aforementioned truncation artifact.

For the pelvis scan, the source-isocenter and the detector-isocenter distances were of 100 cm and 49.9 cm respectively. Only a portion of the available data was used for this paper : a 39.7×4.34 cm^2 detector surface, with 1024×56 rectangular voxels. the FOV was set to 48.75 cm (compared to the original 45 cm) in order to include most of the patient couch. This is necessary to avoid additional artifacts due to objects outside the FOV. The reconstruction matrix was of $416 \times 416 \times 16$ voxels of size $0.117 \times 0.117 \times 0.25$ cm^3. The voxel sizes were consistent with those of the clinical reconstruction. The artifact-free z-span is of 10 slices or 2.5 cm.

In both cases, the z-span is relatively small, limited by the capacity of the GPU's random access memory (RAM) to store the entire projection matrix. Nevertheless, it consists of sub-millimetric slices for the head phantom, so at the cost of a lower resolution in the axial direction, a larger FOV may be reconstructed. The loss of one symmetry for the pelvis scan results in an even lower z-span. The projection matrix was formed using the Siddon's raytracing algorithm [79]. The rays were considered infinitely thin and emitted from a point source, with one ray per detector bin. These hypotheses represent an approximation compared to the problem geometry, since the focal spot and detector pixels both have small but finite sizes. In consequence, for every detector reading, a three-dimensional beam of radiation is attenuated

in the subject. However, as shown by Xu and Mueller [73], the results are similar when using the *line* or *beam* propagation models for simulating typical CT geometries.

The thin-ray model is known to be prone to creating high-frequency artifacts, similar to Moiré patterns, within reconstructions. Using many thin rays per detector bin to form the projection matrix may alleviate this potential problem, at the expense of reducing the sparsity level of the projection matrix. In this work, lower memory usage was preferred to allow for a faster reconstruction, so one ray per detector bin was used. This did not introduce specific high-frequency artifacts in the reconstructions.

For reference FDK reconstructions of simulated data, an in-house code with bilinear interpolation was used for backprojection. A ramp filter weighted by a Hamming window was used on projection data, since this filter is known to offer a good balance between denoising and edge preservation. The FDK reconstruction of clinical data was performed by the manufacturer's software, using the Ram-Lak filter. The choice of filter is based on an actual clinical protocol, which requires a high spatial resolution of bones' boundaries.

2.5.3 Tuning of reconstruction parameters

After setting experimental conditions, it is essential to determine the OSC-TV reconstruction parameters applicable to such a setup. The only previously undocumented parameter is the exponent p for the number of subsets reduction pattern. Fig. 2.3 illustrates different discrete functions for 200 projections and 12 iterations. Nine iterations were necessary to produce an approximately stationary solution in many cases, therefore 12 iterations were selected as a safe number for the reconstructions presented in this paper. The value $p = 1$ corresponds to a linear reduction, $p = 1/2$ to square root and so forth. The relative performance of these patterns was compared using the normalized root-mean-square deviation (NRMSD), defined as follows :

$$\mathrm{NRMSD}(n) \equiv \left(\frac{1}{\mu_{\max} - \mu_{\min}}\right) \sqrt{\frac{\sum_j \left(\mu_j - \mu_j^{(n)}\right)^2}{j_{\max}}}, \tag{2.9}$$

where μ denotes phantom voxel values and $\mu^{(n)}$ the result of the n-th iteration of the reconstruction algorithm. The entire volume free from the truncation artifact was used to evaluate the metric. Different values of p resulted in either a monotonic reduction of the NRMSD or an oscillating pattern. The progression of NRMSD for some representative p values is shown in Fig. 2.4. For $p \in \{1, 1/2\}$, the NRMSD(n) was found to be non-increasing, and for $p \in \{1/4, 1/8, 1/16, 1/32\}$ the metric oscillated during reconstruction. It was also observed that with decreasing values of p, the problem was increasingly numerically ill-posed, resulting in several isolated high-contrast streaks in images. This effect can be explained by the greater image bias induced by high-subsets-number OSC steps for low p values. For higher exponents, namely $p \in \{1, 1/2\}$, there were two to zero low-contrast streaks for the evaluated volume,

consistent with less biased OSC steps. Finally, $p = 1/2$ (square root function) was selected as the default value, since it yields a lower estimation error.

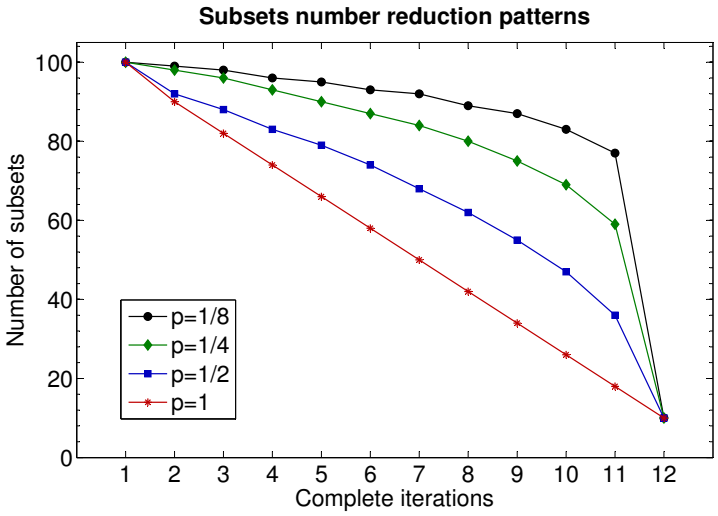


FIGURE 2.3 – Representative subset number reduction patterns for OSC-TV.

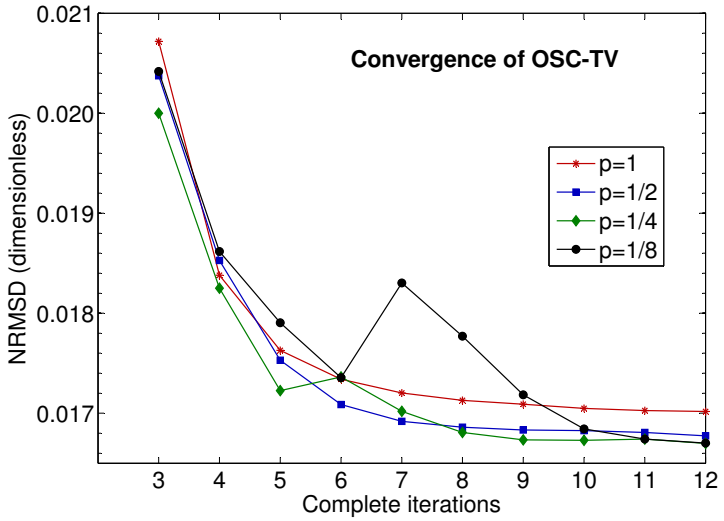


FIGURE 2.4 – NRMSD progression for some representative subset number reduction patterns. For higher p values, the metric is non-increasing ; for lower values, it oscillates.

2.5.4 Convergence of OSC-TV and POCS-TV

Once the reconstruction parameters were fixed, the OSC-TV algorithm was compared to a well-documented TV-regularized method proposed by Sidky *et al.* [56], further denoted as POCS-TV (TV-regularized projections onto convex sets). This method combines alternating ART

and TV-regularization steps and produced good results on noisy and incomplete projection datasets. Numerically, OSC-TV requires larger data structures and more computations per iteration. In consequence, the purpose of the following comparison is to verify if the added cost of performing OSC-TV is beneficial in terms of image quality.

For OSC-TV, the regularization parameter c was empirically determined as 0.2 for the simulated noisy head projections. It was similarly determined for POCS-TV and yielded the same value. The NRMSD was computed for the estimates obtained from 12 consecutive iterations. The execution was extended to 36 iterations for POCS-TV without significant changes in image or its NRMSD, so the result is not shown here. This metric was evaluated on the whole artifact-free volume of 126 reconstructed slices, as well as on a cylindrical region-of-interest (ROI) of 50 pixels in radius and 126 slices in depth. This ROI comprised different anatomical structures, such as air-filled sinuses, skull bones and brain tissues, but no outside air or outer portions of the head, both more prone to streaking artifacts. The NRMSD progressions are shown in Fig. 2.5. Both algorithms show monotonic convergence, however OSC-TV achieves lower estimation error in both cases. This indicates that the additional computational burden of OSC-TV is justified and the objective function of OSC-TV seems more appropriate for noisy X-ray CT data. The next issue to investigate is whether the final results of OSC-TV are satisfactory from an image quality standpoint.

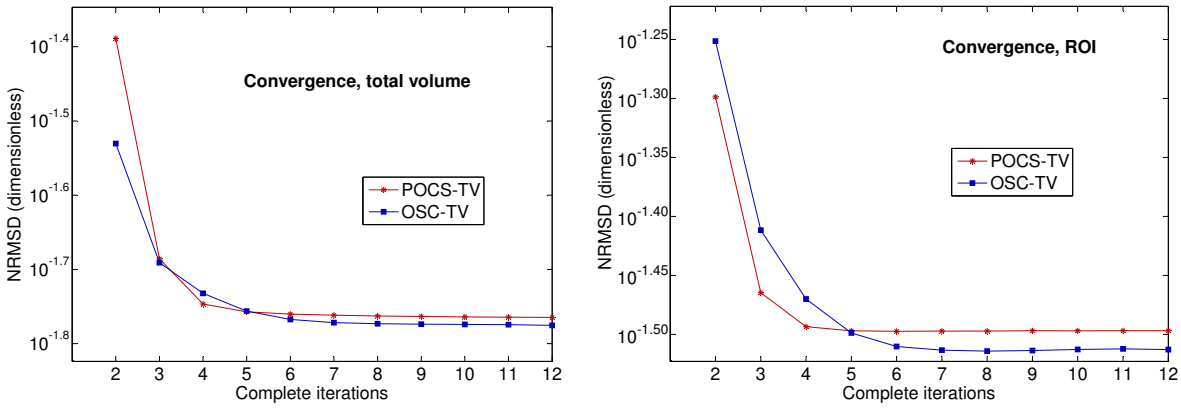


FIGURE 2.5 – NRMSD progression for OSC-TV and POCS-TV, evaluated on the full volume free from the truncation artifact (left) and on a cylindrical ROI (right). In both cases, OSC-TV yields a lower estimation error than POCS-TV.

2.5.5 Image quality

Simulated data

The OSC-TV reconstruction results, with p now fixed to $1/2$, were assessed to show whether the few-view OSC-TV images are of quality comparable to full-dataset FDK filtered backprojection

images. A reference image was reconstructed using FDK on a full dataset of 360 projections, and is used to approximately represent clinical image quality. The few-view projection set was reconstructed using OSC-TV and POCS-TV. Fig. 2.6 shows the reconstruction of the central slice by the three methods. The gray levels in the images are scaled relative to the phantom μ_j values, so 0 (black) is equivalent to the minimum phantom attenuation value and 1 (white) to the maximum. The full dataset FDK reconstruction displays a characteristic radial streaking and image noise. Both effects are not severe, since they are only observable with a narrow-window soft-tissue display setting. The OSC-TV reconstruction yields results similar to the reference image despite an approximately two-fold dose reduction. The structures are relatively uniform and their boundaries are clearly defined. However, the noise affects the shape of the boundaries of the structures. This effect is the major limitation of the TV method, which is not based on the attenuation model but on an image-space regularization criterion. The effects of noise are better perceived when a narrow gray-scale window is used, as can be seen in the second row of Fig. 2.6. POCS-TV performs similarly to OSC-TV, nevertheless, the uniform regions of the phantom are more "patchy" on the reconstruction, a pattern very similar to those shown by Sidky *et al.* [56]

Sagittal plane views of the same reconstructions are shown in Fig. 2.7. As for the axial slices, the full dataset FDK reconstruction yields an acceptable image. The few-view OSC-TV reconstruction effectively eliminates streaking in the axial direction thanks to the 3D TV regularization. In fact, only a few slices at the top and bottom of image are visibly biased. POCS-TV is again free from major artifacts but yields a less uniform result.

Difference between the phantom and reconstructions in both planes of interest are shown as images in Fig. 2.8. When comparing the OSC-TV results specifically to the full-dataset FDK images, the latter show blurry boundaries, consistent with the use of interpolation for backprojection as well as the Hamming window. In contrast, OSC-TV offers a lower noise as well as lower thickness of the boundary errors when compared to full-dataset FDK. POCS-TV behaves similarly to OSC-TV, but the uniform regions and boundaries are slightly noisier in the POCS-TV image.

The contrast-to-noise ratio (CNR) was also measured, 6 slices off-center. The regions of interest were of 20×20 voxels, taken within two areas known to be uniform in the phantom. The results are respectively 33.3, 470.5 and 102.0 and for full-dataset FDK, few-view OSC-TV and few-view POCS-TV. The OSC-TV reconstruction's CNR is significantly higher than in the other cases, which reflects its capacity to suppress noise in uniform regions. However, this metric does not reflect how the organ boundaries are affected by noise. For OSC-TV, variance reduction is favored, while the image suffers from bias. The variance reduction is shown by a high CNR. The bias resides within the "patches" which are locally uniform but differ from the true attenuation values. For POCS-TV, the higher noise in the image seems to be related to its less appropriate objective function, as discussed in section 2.5.4. In fact, extensive tests

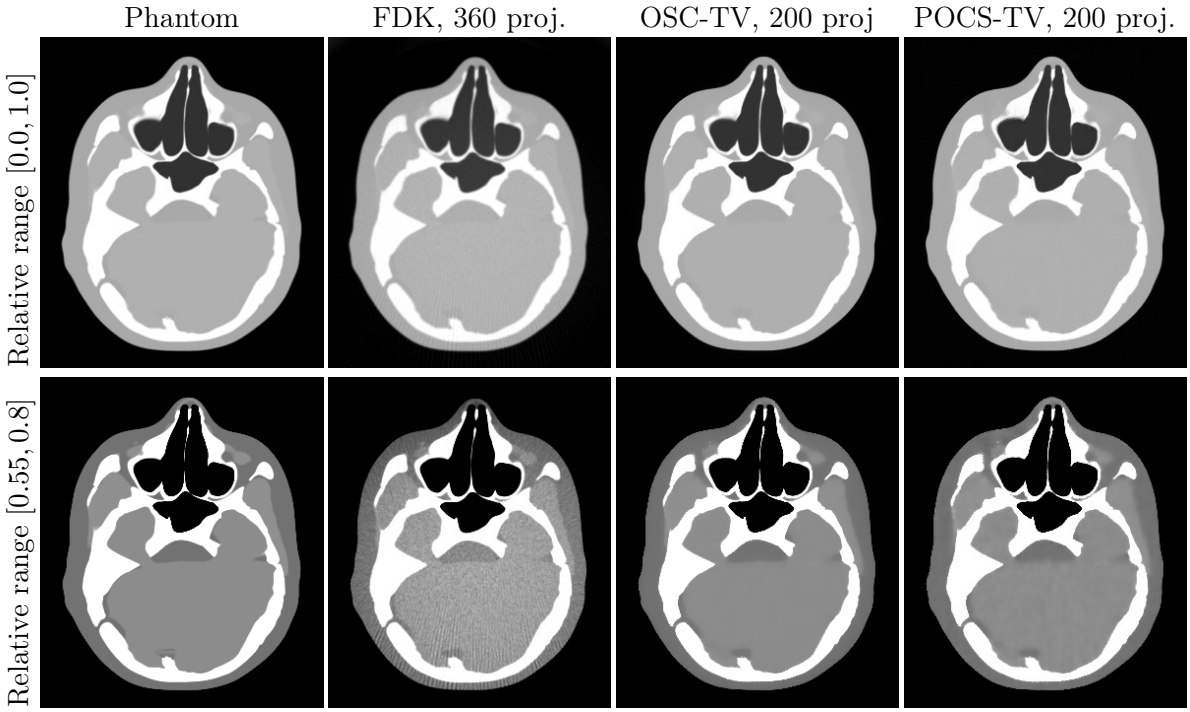


FIGURE 2.6 – Phantom and reconstructions of the noisy projections, central slice. The top row shows the full span of gray values and the bottom row – a narrow window soft-tissue setting. Note the radial streaking and noise for FDK. The noise is mostly suppressed by OSC-TV reconstruction, however structure boundaries are somewhat altered. For POCS-TV, uniform regions of the phantom appear patchy when using a narrow density window. The reconstructions are originally 384×384 voxels and are cropped here to show the region of interest.

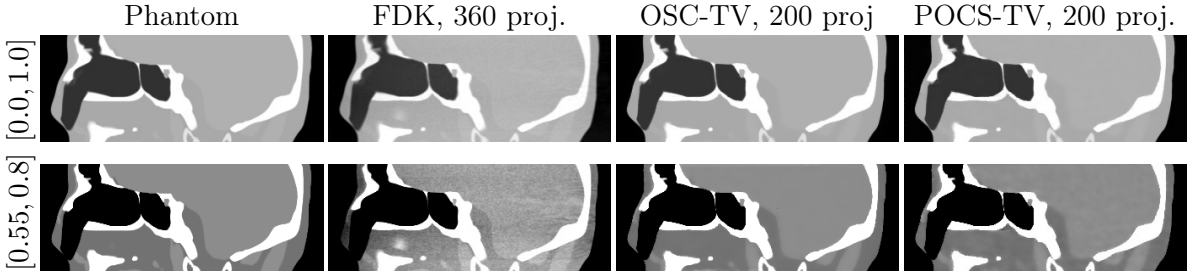


FIGURE 2.7 – Sagittal plane view of the phantom and reconstructions, 126 slices. Note high-frequency artifacts and noise for the FDK reconstruction and residual truncation artifacts for OSC-TV and POCS-TV. The leftmost column indicates relative μ range. The images are cropped to the region of interest.

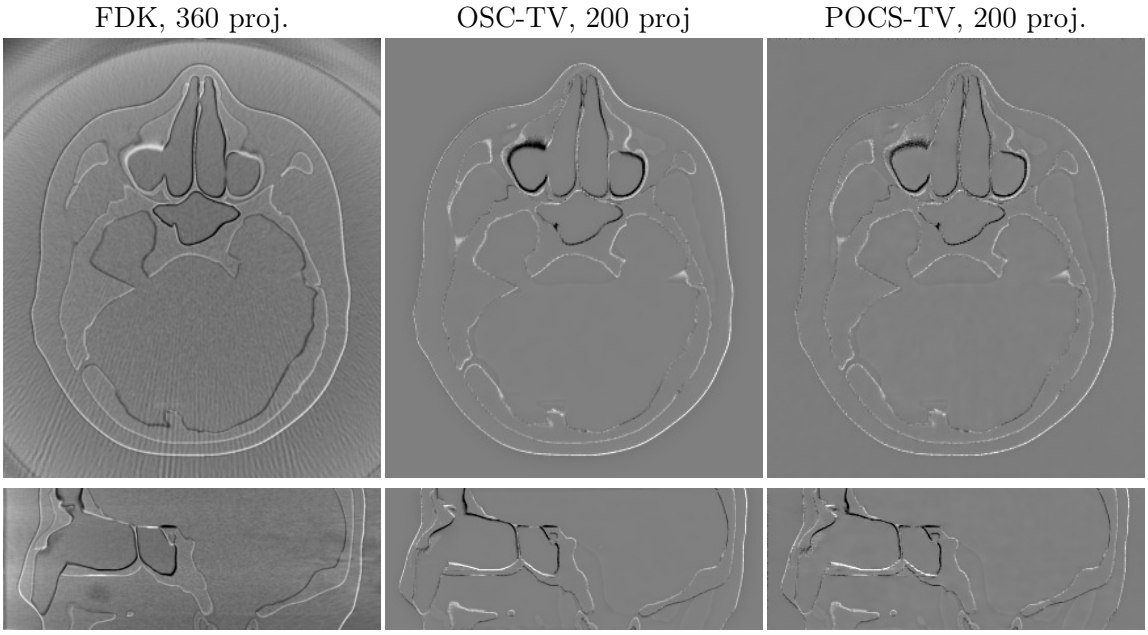


FIGURE 2.8 – The difference between the phantom and its reconstructions, the central axial slice (top row) and a sagittal plane view (bottom row). The relative $\Delta\mu$ range is $[-0.25, 0.25]$. OSC-TV shows a lower noise than FDK, as well as boundaries of lower amplitude and thickness. POCS-TV behaves similarly to OSC-TV. Note that the axial and sagittal slices are not to scale.

with different regularization parameters did not allow for better image quality when using POCS-TV.

Other properties of the OSC-TV algorithm were studied in a 2D reconstruction framework by Matenine [162], namely its modulation transfer function and metal artifact reduction, demonstrating its superiority over filtered backprojection. Overall, the present results suggest that the OSC-TV approach is capable of effectively handling dose reduction, despite a number of algorithmic approximations aimed at accelerating the computations, thus offering a viable alternative to full-dataset FDK as well as few-view POCS-TV.

Clinical data

A pelvic CBCT acquisition was reconstructed by OSC-TV and POCS-TV using the few-view (172 projections dataset). The FDK reconstruction of the full dataset, performed by the manufacturer's software, is included as a clinical reference of image quality and used the Ram-Lak filter. The OSC-TV reconstruction used a low regularization coefficient $c = 0.02$ (cf. 0.2 for the simulated head phantom), while the POCS-TV used $c = 0.1$ to strike a good balance between over-regularization and streaking artifacts. The central slice and coronal projections are presented in Fig. 2.9. The FDK reconstruction used by the manufacturer

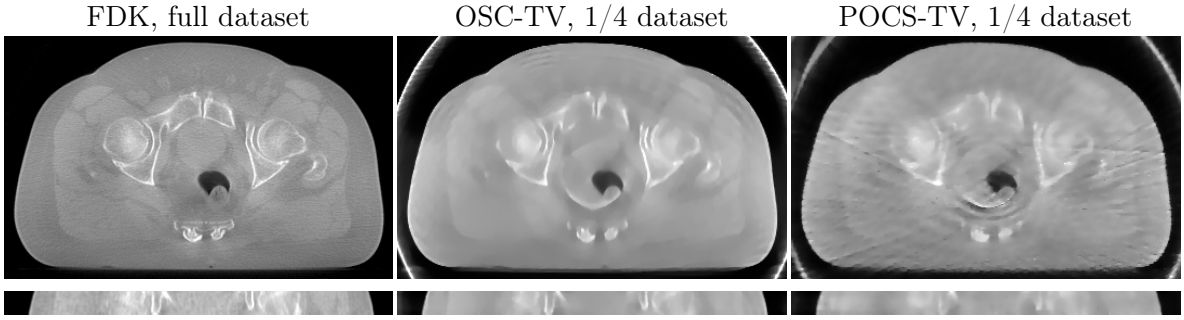


FIGURE 2.9 – Reconstructions of the patient pelvis data. The full span of gray values is shown. The top row shows the central slice and the bottom row – a coronal view at the prostate level. Both OSC-TV and POCS-TV results show two high-density ring artifacts, with an outer ring outside the patient volume and an inner ring at the rectum level. They seem to be due to incomplete projection data pre-processing. The reconstructed slices are originally 384×384 for FDK and 416×416 for OSC-TV and POCS-TV, using the same pixel size for all cases, and are cropped to the region of interest.

includes sophisticated raw data correction not accessible to the authors, so the clinical images are free from gross artifacts. Both OSC-TV and POCS-TV reconstructions are affected by two high-density concentric ring artifacts, a large outer ring and a small ring close to the isocenter. They are most probably caused by simplistic raw data correction, since the projections display abnormal intensity gradients at the boundaries. In addition, the current OSC-TV algorithm does not include beam-hardening nor scatter correction, reducing contrast. The same applies to POCS-TV.

In the OSC-TV reconstruction, some ripple artifacts are visible inside soft tissues, but streaking was mostly eliminated. Compact bone is relatively well defined, so the image could be clinically useful for geometry matching. In the coronal plane, the OSC-TV reconstruction appears to match the full-dataset FDK image in terms of bone structure definition, but lacks contrast as in the axial plane. Overall, despite the limited projection pre-processing, the OSC-TV image renders the main anatomical features. With POCS-TV, evident streaking artifacts are present and could not be eliminated even through careful regularization tuning. In the coronal plane, POCS-TV reconstruction offers few anatomical details, possibly due to over-regularization necessary to mitigate the streaking artifact.

2.5.6 Computation time

In order to validate the applicability of the OSC-TV algorithm in a clinical setting, its current implementation was benchmarked in terms of computational speed. A strict comparison with FDK seems irrelevant, knowing its execution times are in the order of a few seconds [154]. POCS-TV was implemented for GPU using the optimization strategies described in this paper, while the regularization code is identical for both algorithms. Reconstructions were performed

on a NVIDIA Titan GPU, featuring 2688 computing cores with a 1020 MHz maximum clock speed and a 288 GiB/s theoretical maximum memory bandwidth, using 32-bit floating point operations. The total reconstruction time for simulated and clinical acquisitions is shown in Tab. 2.1. The clinical reconstructions are faster, since the projections' z-span is lower, due to the limited size of the projection matrix and the absence of axial reflection symmetry for half-fan acquisitions. A full-fan acquisition with all the symmetries and therefore maximum volume grid resolution requires roughly twice the time for reconstruction, compared to a half-fan geometry. On the other hand, the precomputed projection matrix seriously limits the reconstruction volume for half-fan scans, requiring use of thicker slices. In all cases, the current implementation of POCS-TV processes the same number of projection data samples faster than OSC-TV, however its optimal solution has a higher NRMSD error, as indicated by the simulated results.

TABLE 2.1 – Reconstruction times for different geometries, based on ten repetitions for each entry. The half-fan reconstructions take shorter times due to the reduced detector grid size. The standard deviation is of 2 s or less in all cases.

Volume grid size	Proj.	Detector size	Fan type	$\text{Time}_{\text{OSC-TV}}$ (s)	$\text{Time}_{\text{POCS-TV}}$ (s)
384×384×192	200	512×192	full	124	92
416×416×16	172	1024×56	half	66	54

As mentioned in section 2.4.4, non-negligible time is necessary to read the precomputed projection matrix from disk. Specifically, about 6 s are required to load the matrix for the head geometry and about 5 s for the pelvis geometry. This memory transaction can be executed during clinical data acquisition, which requires about 1 min for current flat-panel CBCT systems. In consequence, it is not included in the total reconstruction time.

It was observed that the 3D TV regularization using the recent NVIDIA Titan GPU was very efficient, requiring a time of roughly 0.4 s per iteration for the synthetic head image and 0.04 s per iteration for the smaller pelvis image. This resulted in total regularization times of about 5 s and 0.5 s per reconstruction, respectively.

These results cannot be directly compared to other GPU implementations of iterative reconstruction, due to the evolution of GPU hardware and differences in experimental protocols. Nevertheless, a cautious comparison to the results of Jia *et al.* [20] is presented. The latter obtained reconstruction times of 130 s when processing 40 projections, also simulating a Varian OBI scanner. The GPU in his case was an NVIDIA Tesla C1060 with 102.4 GiB/s theoretical memory bandwidth. The OSC-TV implementation processes 200 projections using the same time, with a 288 GiB/s theoretical bandwidth. OSC-TV compares favorably, processing 5 times more projections with a GPU offering only 3 times the bandwidth. However, the GPUs' computing performances are 933 GFLOP/s (Tesla) and 4500 GFLOP/s (Titan), so OSC-TV is not necessarily faster than the algorithm by Jia *et al.* [20] In absolute terms,

a 1-2 minutes reconstruction time prevents real-time image visualization. However, it is not expected to hinder a typical clinical workflow, where a patient occupies the room for several minutes. The main computational bottleneck was backprojection, and memory transactions coalescence was essential to achieve relatively fast reconstructions. An acceleration factor of 1.23× for the projection-backprojection kernel was obtained by using the simple subvolume organization of the accumulating matrices, compared to the linearly stored μ data. From a more general perspective, it is worth mentioning that the proposed technique has a drawback common to many iterative algorithms : the entire projection set is required to be available before reconstruction can begin. In contrast, filtered backprojection can be launched during acquisition, which may take about 1 min for flat-panel CBCT systems. Nevertheless, the obtained OSC-TV reconstruction time seems acceptable for clinical non-real-time applications, considering its potential for dose reduction.

2.6 Conclusion

The present study has shown the potential of the combined OSC-TV reconstruction algorithm to reduce CBCT radiation dose by a factor of about two to four, depending on the application. This algorithm delivers accurate reconstructions in presence of noise, while requiring fewer projections, despite algorithmic approximations required to speed up the computations. It appears to be a valuable alternative to the well-known FDK and POCS-TV approaches. Clinical data is more challenging to reconstruct, due to possible loss of symmetries, beam hardening and radiation scatter. The use of recent GPU hardware, combined to easily transferable optimization strategies, yields volume resolutions and execution times compatible with a number of clinical CBCT imaging protocols. Current work focuses on integrating a more sophisticated physics model, including beam-hardening and X-ray scatter, in order to meet the requirements of clinical quantitative CT, as well as further code optimization intended to accommodate wider beams and higher slice resolutions. The GPU memory size is a notable limitation for the current implementation, which pre-computes and loads the entire projection matrix in this memory space. Use of larger memory capacities and multi-GPU implementations may solve this issue. To accommodate arbitrary geometries, an alternative implementation capable of simultaneous raytracing and reconstruction is envisioned.

2.7 Acknowledgements

This work was supported in part by the Fonds de recherche du Québec – Nature et technologies (FRQ-NT). The authors acknowledge partial support by the CREATE Medical Physics Research Training Network grant of the Natural Sciences and Engineering Research Council of Canada (Grant number : 432290). They also wish to thank Julia Mascolo-Fortin for her input regarding the clinical scanning geometry.

Chapitre 3

Evaluation of the OSC-TV Iterative Reconstruction Algorithm for Cone-Beam Optical CT

Dmitri Matenine[1], Julia Mascolo-Fortin[1], Yves Goussard[2] et Philippe Després[1,3,4]

[1] *Département de physique, de génie physique et d'optique, Université Laval, Québec (Québec), Canada*

[2] *Département de génie électrique / Institut de génie biomédical, École Polytechnique de Montréal, Montréal (Québec), Canada*

[3] *Centre de recherche sur le cancer, Université Laval, Québec (Québec), Canada*

[4] *Département de radio-oncologie et Centre de recherche du CHU de Québec (Québec) Canada.*

3.1 Résumé

Cet article présente l'évaluation de l'algorithme de reconstruction itératif OSC-TV dans le domaine de la tomographie optique à faisceau conique. L'un des usages de la tomographie optique est la dosimétrie 3D pour la radiothérapie, où cette modalité est utilisée pour produire une carte de la distribution de dose dans un gel radiosensible. La reconstruction itérative basée sur un modèle de l'atténuation de photons est susceptible d'améliorer la qualité d'image en tomographie optique et d'assurer son expansion dans le domaine de la dosimétrie par gel. L'algorithme OSC-TV a été évalué sur des données expérimentales et certaines simulations numériques complémentaires. L'implantation GPU de cette méthode a permis d'obtenir des temps de calcul comparables à ceux de la rétroprojection filtrée. Les images obtenues via OSC-TV ont été comparées avec les rétroprojections filtrés correspondantes. Des acquisitions et des reconstructions ont été effectuées sur des fantômes de résolution spatiale et d'uniformité, suite

à quoi la fonction de transfert de modulation, l'uniformité et la fidélité des densités ont été évaluées. Les artéfacts dus à la réfraction de la lumière et à la perte totale de signal ont été étudiés également. Les résultats indiquent que OSC-TV surpasse la rétroprojection filtrée en termes de qualité d'image grâce à la modélisation formelle de l'atténuation des photons. On a observé une amélioration de la résolution spatiale et une réduction du bruit, ainsi que la réduction de certains artéfacts. La fidélité de l'estimation des densités est demeurée semblable chez les deux algorithmes et l'artéfact dû aux parois du fantôme n'a pas pu être complètement éliminé avec OSC-TV. En somme, cet algorithme itératif représente une approche intéressante en tomographie optique à faisceau conique, compte tenu de l'amélioration de la qualité d'image et de la réduction des artéfacts.

3.2 Abstract

Purpose : The present work evaluates an iterative reconstruction approach, namely, the ordered subsets convex (OSC) algorithm with regularization via total variation minimization (TV) in the field of cone-beam optical computed tomography (optical CT). One of the uses of optical CT is gel-based 3D dosimetry for radiation therapy, where it is employed to map dose distributions in radiosensitive gels. Model-based iterative reconstruction may improve optical CT image quality and contribute to a wider use of optical CT in clinical gel dosimetry.

Methods : This algorithm was evaluated using experimental data acquired by a cone-beam optical CT system, as well as complementary numerical simulations. A fast GPU implementation of OSC-TV was used to achieve reconstruction times comparable to those of conventional filtered backprojection. Images obtained via OSC-TV were compared with the corresponding filtered backprojections. Spatial resolution and uniformity phantoms were scanned and respective reconstructions were subject to evaluation of the modulation transfer function, image uniformity, and accuracy. The artifacts due to refraction and total signal loss from opaque objects were also studied.

Results : The cone-beam optical CT data reconstructions showed that OSC-TV outperforms filtered backprojection in terms of image quality, thanks to a model-based simulation of the photon attenuation process. It was shown to significantly improve the image spatial resolution and reduce image noise. The accuracy of the estimation of linear attenuation coefficients remained similar to that obtained via filtered backprojection. Certain image artifacts due to opaque objects were reduced. Nevertheless, the common artifact due to the gel container walls could not be eliminated.

Conclusions : The use of iterative reconstruction improves cone-beam optical CT image quality in many ways. The comparisons between OSC-TV and filtered backprojection presented in this paper demonstrate that OSC-TV can potentially improve the rendering of spatial features and reduce cone-beam optical CT artifacts.

keywords : optical CT, iterative reconstruction, cone-beam reconstruction, total variation minimization

3.3 Introduction

Optical computed tomography (optical CT) is a volumetric (3D) imaging technique used to estimate the distribution of linear attenuation coefficients in an optically attenuating or scattering object. One of the applications of optical CT is the reading of dosimetric gels used in radiation therapy. In recent years, the development of new treatment techniques, such as intensity modulated radiation therapy (IMRT) and stereotactic radiosurgery, created a demand for sophisticated 3D dosimetric tools, motivating research in the field of radiosensitive gels [28, 29]. Moreover, gel dosimeters are also promising candidates for application in both low-dose-rate and high-dose-rate brachytherapy [35, 163]. 3D dosimeters are expected to be convenient and reliable when commissioning new treatment delivery techniques or performing quality assurance procedures. The main advantages of dosimetric gels are excellent water-equivalence and exceptional spatial resolution [28]. However, some of the major drawbacks of gel dosimetry are the complexity of the manipulations and in some cases the requirement for a timely reading of the gel. Magnetic resonance imaging (MRI) scan of gel dosimeters is considered a gold standard procedure in terms of image quality [27]. It is relatively costly, time-consuming, and has a lower priority than patient imaging in a clinical setting. In other words, MRI dose readout seems quite challenging for routine clinical dosimetry. Optical CT is an interesting alternative to MRI gel reading and is offered in different designs : laser pencil-beam, parallel-beam, and cone-beam. These designs are characterized by different trade-offs between scanning speed, system complexity, user convenience, and image accuracy. Laser pencil-beam optical CT [25, 29, 164] often requires complex calibration procedures, yields long scan times, and provides a spatial resolution bound by the pencil beam diameter. However, it is less sensitive to stray light, thanks to a narrow angular aperture of the optical signal detector, resulting in more accurate reconstructions. Parallel-beam and cone-beam optical CT both appear as valuable alternatives, reducing scan times by an order of magnitude [25, 164] when compared to optical laser CT. Parallel-beam scanners use complex lens arrangements to offer small light acceptance angles [165], achieving image quality similar to that obtained by laser scanners. Cone-beam optical scanners employ a diffuse light source and a pinhole video camera, yielding a simple design, at the cost of reduced reconstruction accuracy caused by large light acceptance angles. Nevertheless, the shortcomings of this acquisition geometry may be overcome in many aspects by using advanced reconstruction techniques.

Current optical CT systems, regardless of their acquisition geometry, typically rely on various formulations of the filtered backprojection for image reconstruction [40, 166, 167]. In particular, the Feldkamp-Davis-Kress (FDK) [4] algorithm is used for cone-beam projection data. This class of algorithms is not based on the physical model of photon attenuation and therefore

introduces certain image artifacts even using projections with a high sampling rate and signal-to-noise ratio (SNR). Recently, Rankine and Oldham [148] and Doran and Yatigammana [37] demonstrated that iterative reconstruction in optical CT yields improved image quality at the expense of relatively long computation times, with simulations on a 2D reconstruction grid. To our knowledge, iterative reconstruction of experimental data in optical CT has not yet been explored. Therefore, the current paper evaluates reconstructions of experimental data on a 3D grid, and image quality is assessed along with execution times. Iterative reconstruction is computationally intensive and may appear unnecessary in a context where the projection data have a high SNR. Nevertheless, it has the potential to yield more accurate 3D images of optical attenuation coefficient distributions in gels, as shown in the field of x-ray CT [7, 18]. A robust iterative statistical reconstruction approach, the ordered subsets convex algorithm with regularization via total variation minimization (OSC-TV) [168], was used in this work. It incorporates a penalized Poisson log-likelihood objective function that is iteratively optimized via expectation-maximization. Regularization is performed via total variation (TV) minimization [56]. The algorithm has shown interesting artifact-reducing qualities in x-ray CT ; therefore it is relevant to evaluate its properties when reconstructing optical CT data.

3.4 Theory

3.4.1 Forward model

In optical CT, the acquired data are composed of readings of visible light exposure after propagation in a partially attenuating medium. The setup mainly consists of a light source, an attenuating object, and a detector. A set of detector readings acquired in a particular geometric configuration (under a certain angle) is called a projection. To solve an inverse problem like the 3D image reconstruction from 2D projections, an appropriate forward model should be selected. A common forward projection model for visible light rays in a partially attenuating medium is the discretized Beer-Lambert law for monochromatic electromagnetic waves,

$$Y_i = d_i \exp\left[-\sum_j l_{ij}\mu_j\right], \tag{3.1}$$

where i and j are linear indices in the projection data and image vectors, respectively. Y_i denotes the detector photon count after attenuation by the object at reading index i ; d_i is the corresponding unattenuated photon count obtained by illuminating the detector without the imaged object, also known as a blank scan. l_{ij} is a discrete distance element through voxel j and μ_j is the voxel's linear attenuation coefficient. This forward model fits relatively well the basic physics involved in optical CT. In this work, the light source had a narrow emission spectrum, as explained in detail in Sec. 3.5.3. In consequence, μ values are estimated for a single wavelength in this model. Moreover, a monochromatic light source also makes it possible to

assume that the detector is a photon counter and not an energy integrator, thereby fitting the Poisson statistics of the reconstruction method described below. Polychromatic attenuation, optical refraction, and light scatter in turbid media i.e., scatter interactions with microscopic particles in the object, are not modeled by Eq. 3.1. Rather, it is assumed that photons follow straight lines and attenuation is due to photon absorption. This simplified model is expected to be valid when employing refractive index matching liquids and photon-absorbing gels. In other words, the physics of the imaging process may at least partially be conformed to the hypotheses of the reconstruction algorithm. In contrast, potential reconstruction errors due to polychromatic attenuation depend on the type and characteristics of the light source and will be discussed in detail in Sec. 3.6.3 Finally, the reconstruction algorithms which use this forward model are less computationally intensive than ones which take into account polychromatic attenuation, refraction, and scatter.

3.4.2 OSC-TV reconstruction algorithm

Iterative reconstruction has been shown to significantly improve image quality in x-ray CT [7, 18], and, considering the similar forward models, is expected to demonstrate similar behavior in optical CT. For this work, OSC-TV has been selected among numerous iterative methods because of its ability to reduce noise and mitigate streaking artifacts in x-ray CT, and because it depends on very few free parameters [168]. It is worth outlining the mathematical formulation and some intrinsic properties of the OSC-TV algorithm in order to highlight its potential usefulness in cone-beam optical CT.

Let t_i denote the *total attenuation* along a ray : $t_i \equiv \sum_j l_{ij}\mu_j$ from Eq. 3.1. Each detector reading Y_i is assumed to follow a Poisson distribution, since each photon leaving the source is subject to the Beer-Lambert law on the ray path, with the probability $\exp(-t_i)$ to reach the detector. The $\{Y_i; \forall i\}$ are assumed to be independent from each other and Y is a random vector representing the entire projection dataset. The μ_j act as distribution parameters and the maximum likelihood method is used to perform the estimation of these parameters. Using this framework, the Poisson log-likelihood objective function was first proposed by Lange and Carson [51] for application in CT. Its full expression is the following :

$$L(\mathrm{Y}, \mu) \quad = \quad \sum_i \{-d_i \exp(-t_i) - Y_i t_i + Y_i \ln d_i - \ln Y_i!\}. \tag{3.2}$$

Considering that the detector photon counts Y_i (after attenuation) and d_i (before attenuation) are known and constant as a result of a CT acquisition, only the terms which depend on μ_j are necessary to maximize the objective function. Hence, its reduced form is [52, 53]

$$L(\mathrm{Y}, \mu) = -\sum_i (d_i e^{-t_i} + Y_i t_i), \tag{3.3}$$

The maximum is obtained via the ordered subsets convex (OSC) expectation-maximization algorithm [153], modified to properly integrate TV regularization [168]. OSC incorporates a

non-negativity constraint, allowing for more accurate μ values on structure boundaries. The Poisson statistics properly model low photon counts, which allows for a model-based reduction of streaking artifacts due to photon starvation. In the current implementation, the elements of distance l_{ij} are obtained via the Siddon's raytracing algorithm [79].

Regularization is performed separately after each OSC iteration, using the total variation minimization approach [56]. The TV regularization is not based on a physical model, but instead on an image-space regularity criterion. The algorithm minimizes the resulting image total variation, which is defined as the sum of the magnitudes of the discrete gradients computed at each voxel of the image. For a 2D image, the total variation is

$$\begin{aligned} ||\mu||_{TV} &\equiv \sum_{a,b} |\vec{\nabla}\mu_{a,b}| \\ &= \sum_{a,b} \sqrt{(\mu_{a,b} - \mu_{a-1,b})^2 + (\mu_{a,b} - \mu_{a,b-1})^2}, \end{aligned} \tag{3.4}$$

where a, b are respectively the column and row pixel indices of the image. The extension of the formalism to a 3D image is straightforward [56, 162]. This criterion is minimized via a gradient descent, for each image estimate resulting from an OSC iteration. This method is efficient for image denoising, assuming noise of high spatial frequency ; therefore it appears very appropriate for optical CT image regularization. Unlike x-ray CT, optical CT does not involve ionizing radiation ; hence, it does not impose imaging dose limits. Therefore, statistical noise becomes apparent only when almost opaque structures are present in the field of view (FOV). Nevertheless, the optical imaging process leads to high-frequency artifacts induced by scratches, floating debris, convection currents etc., as explained in Sec. 3.5.3. TV regularization reduces such artifacts, improving image uniformity. The trade-off of this method is the potential over-regularization of image features of very small physical size and low relative contrast, see Sec. 3.6.5.

TV regularization is known to perform best when the reconstruction step (OSC in this case) induces less and less corrections to the image as iterations progress [56]. In consequence, a gradual subset number reduction technique was used [168] to ensure optimal performance of the TV regularization. This algorithm has a relatively high computational cost and was implemented on commodity graphics hardware (GPU) using the CUDA architecture [92] to yield reconstruction times of about one to two minutes. In comparison, it is common to consider that the total time to obtain a dose distribution image from a gel should be under one hour to be competitive with MRI scanning [169].

3.5 Materials and Methods

3.5.1 Phantoms

Optical CT devices are usually employed to read two main types of radiosensitive substances : radiochromic gels and polymerization gels. Both types of gels are near-transparent before irradiation. In radiochromic gels, the absorbed dose increases the linear attenuation coefficient of particular wavelengths of visible light. For example, ferrous xylenol orange turns from pale orange to purple after absorbing ionizing radiation. Polymerization gels are formulated in such a way that ionizing radiation induces polymerization of monomers present in the gel ; the formed aggregates have a light scatter coefficient related to the absorbed dose [35]. These gels have been shown to be poorly imaged by cone-beam optical CT scanners, because of a significant cupping artifact induced by scattered light [40]. It should be noted that the gelatin used to spatially fix radiochromic gels also causes non-negligible scatter ; however, its behavior can be relatively easily taken into account [41]. The current work focuses on the reading of phantoms which absorb light similarly to radiochromic gels, since they are better suited for cone-beam optical CT scanning.

This paper compares the image quality achieved using the analytic FDK and the iterative OSC-TV reconstruction algorithms for cone-beam optical CT scanning. Evaluation of spatial resolution, the corresponding noise levels, image uniformity, and accuracy was performed. The spatial resolution was measured using the modulation transfer function (MTF) computed via a profile of the edge response function (ERF) [170], which conveniently yields the entire MTF curve using one experimental profile. Reconstruction uniformity and accuracy were measured using μ profiles in uniform media. Qualitative evaluation compared the extent of reconstruction artifacts, such as concentric rings and irregularities surrounding opaque objects.

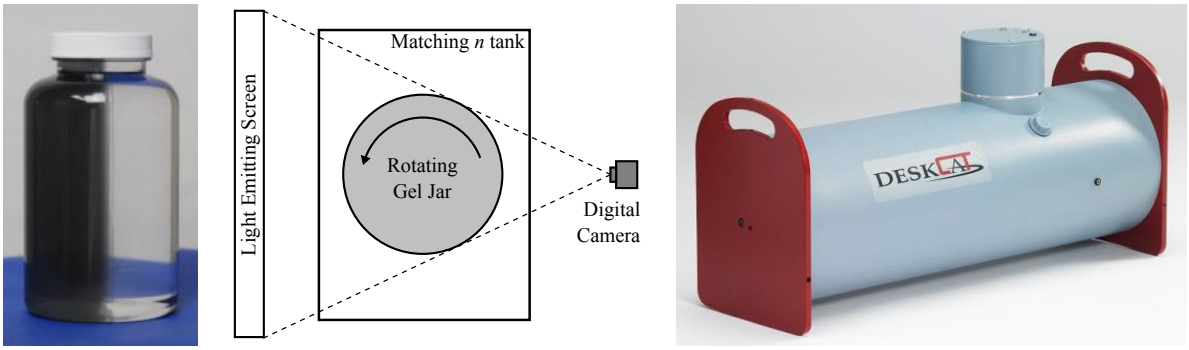
The comparative image quality analysis was completed using different optical phantoms bundled with the DeskCAT™ educational cone-beam optical CT system, manufactured by Modus Medical Devices, Inc. (London, ON, Canada). They consisted of clear plastic jars of 7.2 cm in diameter. A jar containing two gels with different μ values, forming a sharp edge, was used for the MTF evaluation and is shown in Fig. 3.1(a). To evaluate the artifacts induced by opaque objects, a special line-pairs phantom was used. It consisted of an air-filled jar, containing a clear plastic sheet with printed line patterns. A mouselike phantom formed by gels of different μ values was supplied by the scanner manufacturer and, was used for qualitative evaluation of 3D reconstruction artifacts.

Smaller jars (6.8 cm in diameter) filled with water and different volumes of added dye solution were used to evaluate the spatial uniformity of the system response as well as the accuracy of the estimate of μ. The dye (food coloring) was provided by the manufacturer. A reference sample contained water. Five attenuating samples were obtained by adding the following total quantities of dye solution : (2.5±0.5) ml, (5.0±0.5) ml, (7.5±0.7) ml, (10.0±0.7) ml

and (12.5±0.9) ml. Each sample was thoroughly stirred, checked for air bubbles, read by a transmission spectrophotometer, and finally scanned by the DeskCAT device.

3.5.2 Spectrophotometers

A modular registering spectrophotometer (product name : *RS System*) by Laser Components (Chelmsford, UK) was used in order to obtain reference μ values for accuracy evaluation. Its beam diameter, measured at the center of the sample compartment, was equal to (2±1) mm. The dyed water samples were scanned to obtain absorbance measurements using 2 nm steps. The sample had to be aligned manually along the light beam in the sample compartment, for the near-transparent and dyed jars, which induced a total 0.25 % error on the resulting μ values. An emission spectrometer, QE65PRO by Ocean Optics (Dunedin, FL), was used to measure the emission spectrum of the scanner light source, using roughly 0.75 nm steps, to locate the DeskCAT's light emitting diodes (LEDs) emission peak and to properly simulate reference polychromatic attenuation. It was coupled to the light source using a sheathed optical fiber.



(a) (b) (c)

FIGURE 3.1 – (a) Typical gel jar, edge phantom shown. (b) Simplified cone-beam optical CT geometry. Note that the photons' paths are in the opposite direction compared to x-ray tomography ; figure not to scale. (c) The scanner module of DeskCAT™ system. Images (a) and (c) provided by Modus Medical Devices, Inc.

3.5.3 Optical cone-beam scanner

A schematic view of the acquisition geometry and the DeskCAT scanner used in this work are shown in Fig. 3.1(b) and 3.1(c), respectively. This system is designed for educational purposes and mimics an x-ray CT scanner in many aspects. It comprises a diffuse light source, which is a screen illuminated by LEDs with a narrow emission spectrum. The device features red and green LEDs and the corresponding color selector. The incident light crosses a glass window and a tank filled with a liquid in which the gel jar is submerged. The liquid, commonly referred to as *refractive index matching liquid*, has a refractive index very similar to that of the phantom, so

the entire system behaves like a window and not like a lens [25]. The phantom is rotated using a stepper motor, while the light source, tank, and detector remain static. The light rays cross a glass exit window and reach a CMOS video camera that captures projections. The camera exposure time and detector gain are adjustable so that the full dynamic range of the detector is used. This system mimics a cone-beam x-ray CT geometry, where the camera's focal point is equivalent to a point source and the screen is equivalent to a flat panel detector, except that the rays' propagation direction is reversed. This acquisition geometry provides potential advantages, namely, low costs and short scan times, but some challenges remain. For instance, refraction is not totally eliminated using the matching liquid, since the refractive indices of the phantom and the liquid are not perfectly matched. Convection within the matching liquid provokes local variations of refractive index, thereby inducing high-frequency artifacts [25]. Moving debris in suspension create undesired attenuation, which also leads to artifacts, since debris are located outside the reconstructed FOV. Light rays are scattered within the gel phantom, because the gelatin or its substitute is a colloid and not a true solution [171]. All these phenomena are difficult to simulate rigorously, although some progress has been made to handle refraction [37]. We conjecture that a regularized iterative algorithm can improve optical CT image quality using *a priori* knowledge about the scanning system and image properties.

Before acquiring attenuated readings Y_i, a scan of a nearly transparent phantom, or blank scan, is necessary to measure the unattenuated signal d_i. This procedure takes into account the light attenuation in the scanner's windows and matching liquid, as well as the attenuation due to the phantom jar wall. Some of the basic sources of image irregularities are also handled : potential camera vignetting (loss of signal at the image periphery due to camera optics) and the differences of sensitivity in individual camera pixels.

3.5.4 Acquisition and reconstruction parameters

For all of the results presented below, the acquisitions were composed of 400 projections, as well as their respective blank scans. In addition to these raw data scans, calibration All the acquisitions were performed using red LEDs. The native camera resolution was 640×480 pixels, and projection images were cropped according to the 3D volume coverage. Each projection acquired by the DeskCAT's camera is encoded as a 16-bit unsigned integer bitmap. The physical distance from the isocenter to the camera chip is about 40 cm, and the exact focal distance is found by numerical calibration, done for each combination of transparent phantom and matching liquid. The diffuse light source is assumed to be the glass window on the LED panel side and is situated at 4.5 cm from the isocenter. It should be noted that this window is fixed relative to the isocenter and limits the maximum diameter of a cylindrical phantom to 8 cm. The same window also physically limits the covered cylinder height to 7.5 cm. The slice resolution was set to 320×320 cubic voxels with 0.25 mm sides, except for the mouselike phantom, where it was set to 192×192 cubic voxels with 0.417 mm sides. The axial size of the

reconstruction was of 64 slices for all the reconstructed volumes. The total volume coverage was 8.0×8.0×1.60 cm^3 for the high-resolution reconstructions and 8.0×8.0×2.67 cm^3 for the mouselike phantom reconstruction.

3.5.5 Computational aspects of the reconstruction

OSC-TV reconstruction was performed on an NVIDIA® (Santa-Clara, CA) Titan™ GPU, fitted with 2688 computing cores distributed among 14 multiprocessors. The global random access memory (RAM) size was 6 GiB for the GPU. The current implementation of OSC-TV relies on a precomputed and compressed system matrix, so the reconstruction grid size is bound by the size of GPU RAM [168], limiting the z-axis spans of the images to 64 slices in all cases.

By design of the OSC-TV algorithm, OSC reconstruction and TV regularization steps alternate. An OSC iteration consists of a projection through the image estimate and a backprojection of small correcting values. The projection and backprojection are integrated into a single CUDA *kernel* (a single code routine executed by many threads for different data). A single kernel launch computes a certain number of ray projections and backprojects along these same rays. The groups of rays are such that no correcting values are written concurrently, thus avoiding race conditions. The OSC kernel is divided into *thread blocks*, where a block typically performs computations for two rays. In turn, the number of threads per block depends on the number of voxels crossed by the two rays. The block sizes are predetermined as powers of two, so idle threads are present for rays of arbitrary length. In return, the code avoids complex conditional statements. Shared memory is used to store repeatedly requested data, such as the l_{ij} vectors and the references to crossed voxels. For the TV regularization step, the volume is divided into *subvolumes* with a typical size of 16×16×4 voxels and blocks of the same size are launched to compute the gradient of the total variation. For this step, some threads are idle during the upload of the subvolume neighborhood into shared memory. All simple vector operations were implemented using the CUDA Thrust code library [100].

Reference FDK reconstructions were obtained using an in-house CPU code, executed on an Intel® Core™ i7-960 CPU with a 3.20 GHz clock (one core used). The algorithm was set to use a ramp filter multiplied by a Hamming window. It should also be noted that FDK images were originally output in arbitrary units. In order to obtain μ values, a scaling factor was obtained for each FDK image. The raw FDK image was reprojected via the Siddon's raytracing algorithm, for four equally spaced projections. The sum of squares of the differences between the reprojected and the experimental total attenuation values t_i was minimized to obtain a scaling factor for the raw image values. In contrast, model-based OSC-TV reconstruction intrinsically yields estimates of the absolute values of μ, with no need to evaluate a scaling factor.

3.6 Results and Discussion

3.6.1 Edge phantom analysis

The edge phantom scan was reconstructed using FDK and OSC-TV. The central slice reconstructions are shown in Fig. 3.2. The display interval of $[0.0, 0.5]$ cm^{-1} for μ corresponds to the two plateau values of the edge phantom, so the edge response is well observed. The FDK image features multiple concentric ring artifacts. The OSC-TV algorithm produces more uniform images, without the rings. The coronal views of the reconstructed volume confirm the noise-reducing properties of OSC-TV.

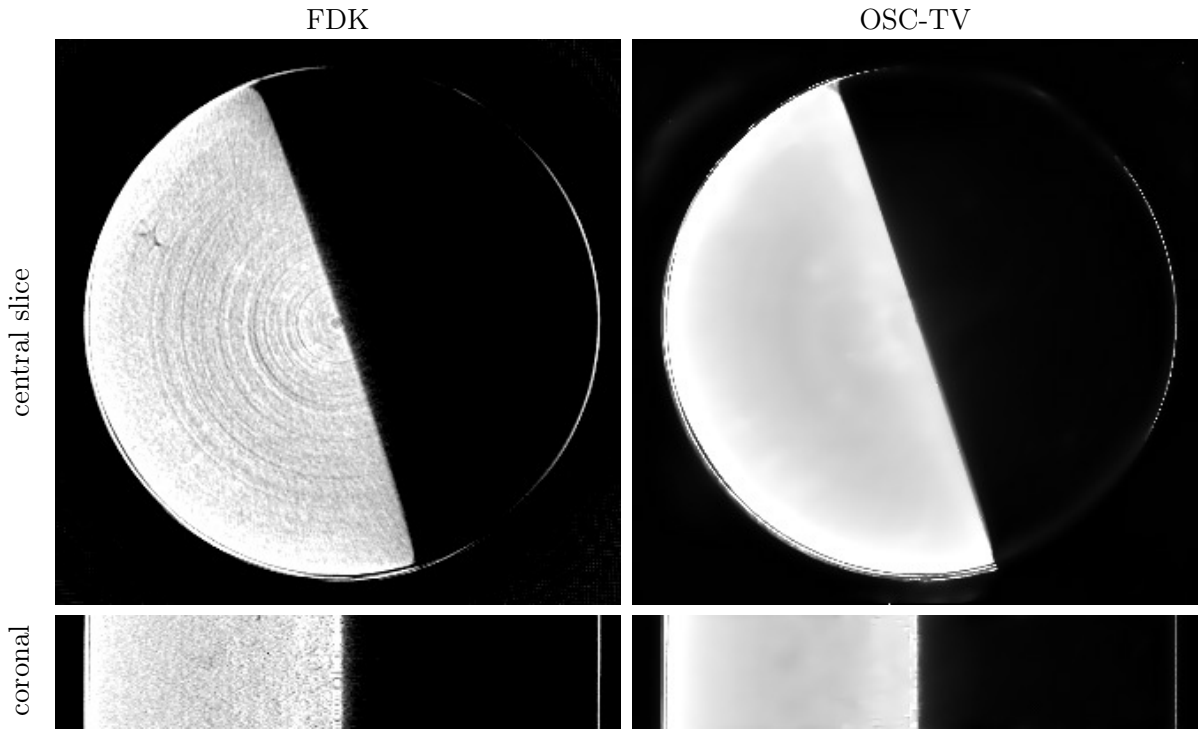


FIGURE 3.2 – Reconstructions of the edge phantom, central slice, and coronal view crossing the isocenter. The grid size is 320×320×64 and the voxel size is 0.25 mm (isometric), the μ interval is $[0.0, 0.5]$ cm^{-1}. Note the ring artifacts in the FDK image, absent in OSC-TV images.

In order to compare spatial resolution, the ERF was measured for the FDK and OSC-TV reconstructions, within the central slice. The profile was acquired perpendicularly to the visible edge, at a distance of 1.8 cm from the isocenter, with a sampling width of 3 voxels. The profiles are shown in Fig. 3.3. The FDK profile is noisier, and its lower plateau has a positive slope. In contrast, the OSC-TV profile is clearly smoother and its upper plateau has a negative slope. This latter effect seems to be due to cupping within the optically dense region, likely due to

scattered light. The following hyperbolic function [172] was fitted to the profiles :

$$\mu(x) = \frac{\mu_{\max} \sinh(\xi a)}{(\cosh(\xi(x - x_0)) + \cosh(\xi a))} + \mu_{\min}, \tag{3.5}$$

where $\mu_{\max}$ and $\mu_{\min}$ are the extreme values of the profile, ξ characterizes the edge gradient, x_0 allows for horizontal translation, and a is the width of the quasirectangular function. This function was selected because it accurately models penumbrae [172]. To satisfy its symmetric shape, the edge dataset was extended via symmetry along the position axis in order to resemble a quasi-rectangular function with penumbrae at the edges. The resulting fits, where only the increasing edge is shown (Fig. 3.3), reproduce well the main gradient, while the irregularities within the plateaus are not reproduced.

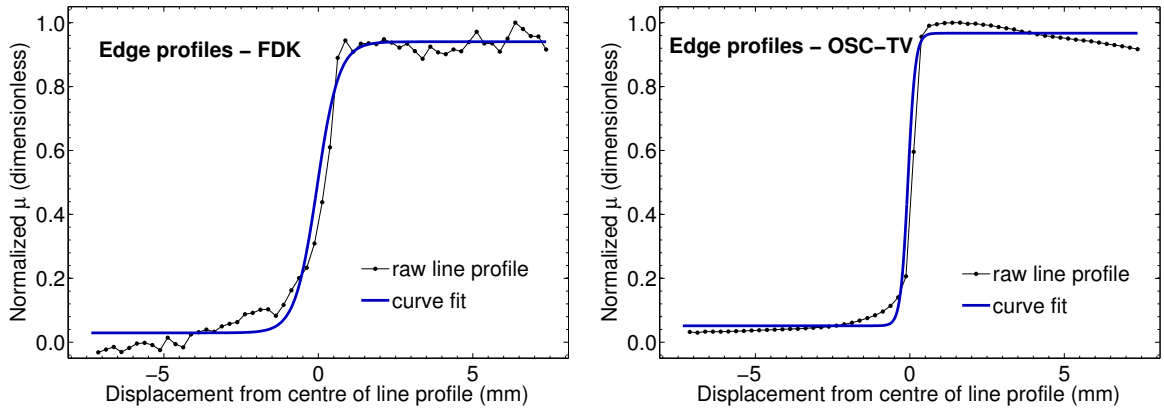
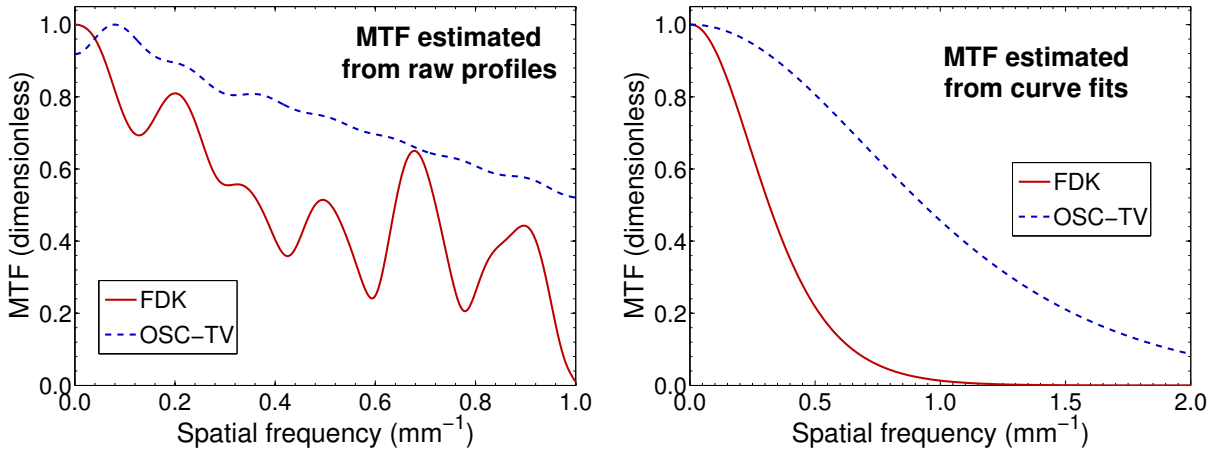


FIGURE 3.3 – Edge response functions fitted to experimental profiles, for (a) the FDK algorithm and (b) the OSC-TV method. Note the Mach band at the upper edge of the OSC-TV reconstruction. The FDK profiles are noisier in the supposedly uniform regions.

The MTF profiles based on the raw and fitted ERFs are shown in Fig. 3.4. Based on the raw profiles, the MTF extracted from the OSC-TV image displays an atypical shape at very low frequencies, so the maximum response is reached at 0.08 mm^{-1}. The MTF then steadily decreases to reach 0.52 at 1 mm^{-1} spatial frequency. The MTF extracted from the FDK image shows more local variations due to noise but is mostly below the MTF of OSC-TV, reaching zero at 1 mm^{-1}. For the fitted profiles, the estimated spatial resolution is again better for OSC-TV.

The above MTF estimations were complemented with measurements of image noise. Since the filtering techniques of the algorithms differ, it is possible, at least in theory, that the higher MTF is obtained by OSC-TV at the price of a higher image noise. Square regions of interest (ROIs) of 32×32 (1024 voxels total) were taken within the studied slices, one in the dark (optically transparent) area and one in the bright (optically attenuating) area within the phantom image. The ROI standard deviation σ was computed, and the resulting values are



(a) (b)

FIGURE 3.4 – MTF profiles based on (a) raw and (b) fitted ERFs. In both graphs, the MTF obtained from the OSC-TV image is higher at higher frequencies when compared to FDK, suggesting a better spatial resolution.

shown in Tab. 3.1. In the FDK image, it was 5.8 times higher than in the OSC-TV image for the dark area and 8.5 times higher for the bright area. In sum, OSC-TV simultaneously offered noise reduction and improvement of spatial resolution. The evaluation of MTF and

TABLE 3.1 – Standard deviation within uniform areas of the edge phantom. Note the lower values for OSC-TV vs FDK.

Sample	Dark area σ (cm^{-1})	Bright area σ (cm^{-1})
FDK	0.013	0.033
OSC-TV	0.0023	0.0039
σ ratio	5.8	8.5

noise levels was also performed 20 slices (5 mm) off-center and yielded very similar results, not presented here for brevity.

3.6.2 Opaque line reconstruction

The presence of opaque objects leads to the absorption of all the incident photons and a complete loss of signal for certain attenuation paths. Total signal loss is known to cause severe artifacts in FDK reconstructions. This phenomenon is not well modeled by the Beer-Lambert law either, since a nonzero signal is expected for any path of finite length and linear attenuation coefficient. However, the positivity constraint on μ integrated into this forward model reduces the extent of the artifacts. Fig. 3.5 shows a single slice of the reconstruction of the line pairs phantom, containing a single line. One can note that the OSC-TV algorithm removes the negative *black holes* around the ends of the line and reduces the *halo* surrounding the edges

of the line. This latter halo effect was also analyzed quantitatively. *In-slice* halo profiles were

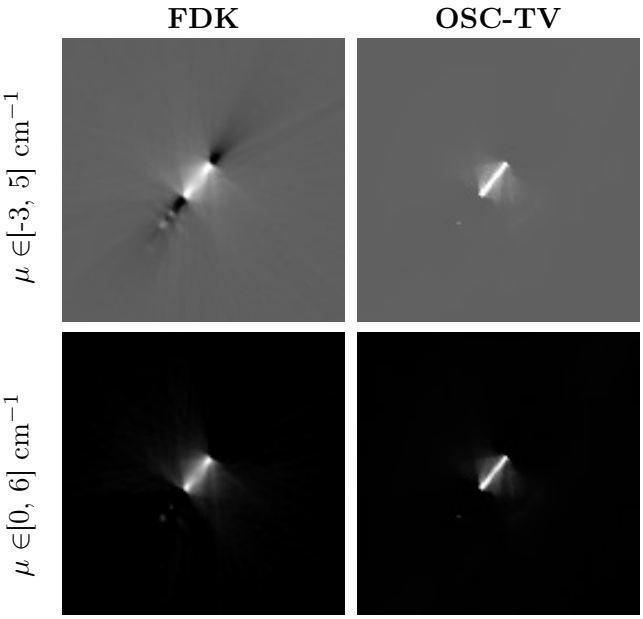
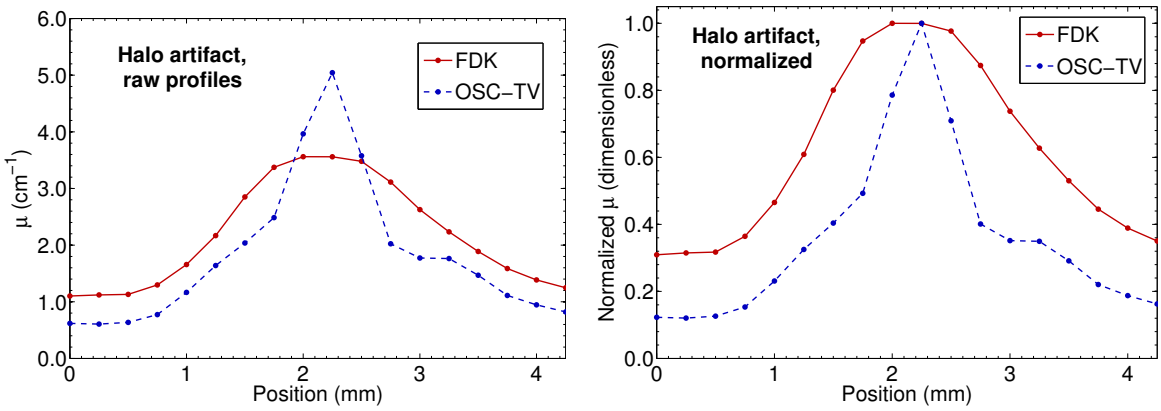


FIGURE 3.5 – A single opaque line reconstruction, central slice, cropped to 128×128 voxels. Upper row : μ interval set to show the *black hole* artifact. It exposes streaking and the negative μ values produced by the FDK algorithm and absent from the OSC-TV reconstruction. In the lower row, note the halo surrounding the line, more prominent for FDK than for OSC-TV.

acquired from the slices shown in Fig. 3.5, perpendicularly to the opaque line, crossing the middle of the line. In terms of absolute μ, OSC-TV yields a higher peak, more representative of an opaque object (5.0 cm^{-1} for OSC-TV vs 3.6 cm^{-1} for FDK), see Fig. 3.6(a). It also yields a significantly narrower profile, with a full width at half-maximum (FWHM) of 0.9 mm for OSC-TV, compared to FWHM=2.5 mm for FDK, demonstrating effective artifact reduction, see Fig. 3.6(b).

3.6.3 Image uniformity and accuracy

Uniformity analysis was performed on dyed water jars. Reconstructed central slices for the water phantom with 5.0 ml added dye are shown in Fig. 3.7(a) and 3.7b). The FDK reconstruction features concentric circular artifacts. These are attenuated with OSC-TV. Both reconstructions display high-density outer rings, likely due to refraction at the jar wall, as well as misalignment between the near-transparent phantom scan and the dyed phantom scan. This boundary artifact is less prominent with OSC-TV. Uniformity was further studied via profiles taken in the central slice along the x-axis, see Fig. 3.7(c). The profiles do not cover the entire jar diameter to avoid displaying the peaks due to jar walls, which vary by an order of magnitude. For 5.0 ml dye, OSC-TV yields a uniform profile and its falloff at the edges is slower than for FDK. The latter seems to be affected by a bowing artifact. For 10.0 ml dye,



(a) (b)

FIGURE 3.6 – Halo profiles for the opaque line phantom, (a) absolute μ values and (b) normalized profiles. Note the higher peak value for OSC-TV (5.0 cm^{-1} vs 3.6 cm^{-1}) ; also a narrower peak for OSC-TV (FWHM=0.9 mm for OSC-TV vs FWHM=2.5 mm for FDK).

the OSC-TV profile shows a cupping artifact (about 8% contrast loss at center), while FDK is still affected by bowing. The OSC-TV algorithm itself does not feature signal correction functions which would explicitly mitigate a potential bowing artifact or cause a cupping artifact. Hence, the cupping observed for the OSC-TV profiles is likely to have a physical cause. In the literature, this phenomenon is attributed to light scatter due to colloidal additives, see Olding *et al.* [40], and is largely taken into account when the reference scan of the nearly transparent jar is performed. In this work, pure water was used as matching liquid and dye solvent. Therefore, scatter and refraction were kept to a minimum and were unlikely to cause cupping. Alternatively, the polychromaticity of the light source is a likely source of the cupping. The dye absorption spectrum is invariable across the jar ; a sample spectrum is shown in Fig. 3.8 (a). However, the light beam spectrum varies across the jar in a manner similar to beam hardening in x-ray CT. A more appropriate term would be *spectrum flattening*, since both "tails" of the beam spectrum become more prominent. A shift of the peak toward the red also occurs. The plots of the simulated spectrum flattening in the uniform phantoms are provided in Appendix A. In consequence, the light beams crossing the center of the jar are more intense than expected by a monochromatic algorithm, and underestimation of μ at the center of the volume occurs. The conjecture regarding non-negligible spectral effects is further supported by the accuracy analysis developed further in the section.

The bowing in FDK images was investigated as well, using a discrete numerical phantom fully mimicking a uniformly attenuating cylindrical phantom. The projections were acquired using Siddon raytracing. Significant bowing was observed in the FDK reconstructions of the synthetic projection data : a 10% contrast loss at the cylinder edges and a 4.0% contrast gain in the middle of the profile. Such bowing would subsist even for the jar with 10.0 ml

added dye, where OSC-TV results suggest a cupping artifact. The reconstruction was also performed using filtered backprojection with the weights provided by the Siddon method (l_{ij}) and the same bowing artifact was observed. A similar finding was reported by Jaffray and Siewerdsen [173] regarding experimental and simulated flat-panel cone-beam CT data. In consequence, this artifact seems to be intrinsic to the flat-panel geometry and requires specific corrections, which are out of the scope of this paper.

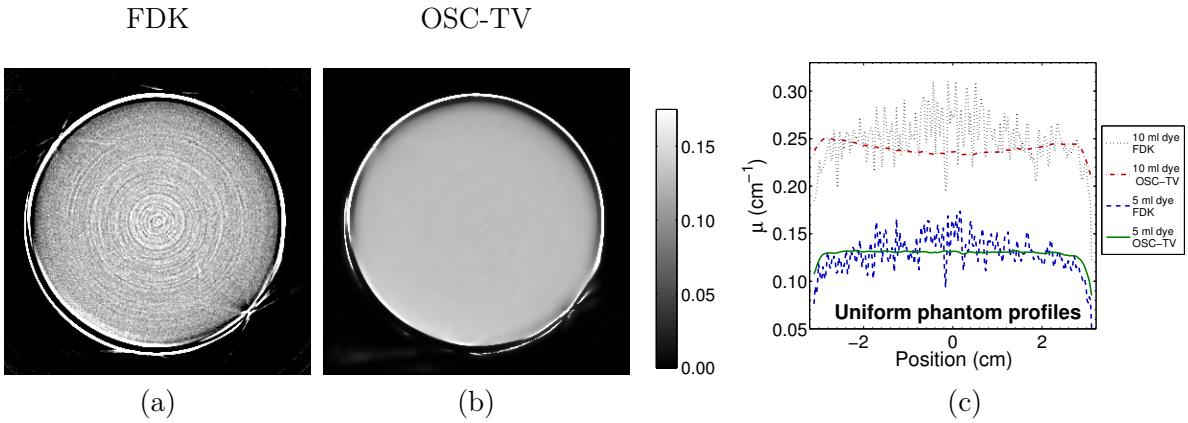


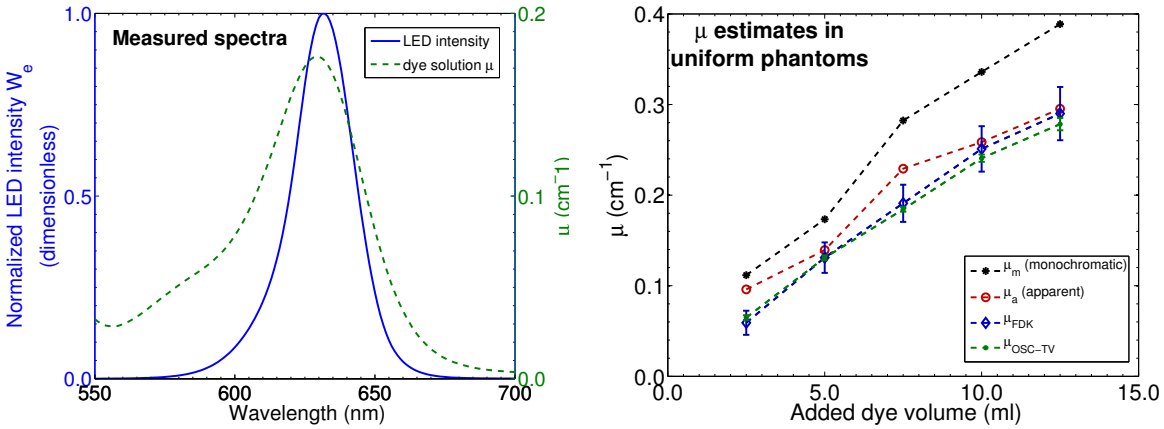
FIGURE 3.7 – Uniform phantom analysis. Central slices of the uniform 5.0 ml dye phantom reconstruction (a) FDK and (b) OSC-TV, μ interval [0.0, 0.175] cm^{-1}. Note the circular artifacts for FDK. (c) Central slice profiles for 5.0 ml and 10.0 ml dye phantom. For OSC-TV, note effective denoising in all cases and non-negligible cupping at 10.0 ml dye. For FDK, note slight bowing.

Imaging accuracy evaluation was performed on uniformly dyed water jars, by comparing reconstructed μ values with ones independently acquired using spectrophotometers. Let λ represent light wavelengths. First, the emission spectrum of the DeskCAT red LEDs, denoted $W_e(\lambda)$, was measured. It provided the exact wavelength of the emission peak, which was of 632 nm for the machine used. The relevant portion of the emission spectrum (550-700 nm) is shown in Fig. 3.8(a).

Second, a collimated spectrophotometer was used to acquire absorbance measurements, using a water-filled jar as the "nearly-transparent" sample. Absorbance measurements were converted to optical density (OD) and then, using the jar diameter, to $\mu(\lambda)$ values. A representative absorption spectrum (550-700 nm interval for 5.0 ml added dye) is shown in Fig. 3.8(a).

Knowing the peak emission wavelength, monochromatic attenuation coefficients were drawn from the jars' absorption spectra. For each jar, μ values measured for 630, 632 and 634 nm were averaged and designated as the monochromatic μ_m. The latter was computed for each jar and taken as a gold standard for the accuracy evaluation, see Fig. 3.8(b).

These measurements allow one to simulate the polychromatic attenuation within each sample



(a) (b)

FIGURE 3.8 – Image accuracy analysis. (a) Measured spectra of the DeskCAT red LEDs and a sample dye attenuation (for 5.0 ml added dye) in the relevant region. (b) μ estimated by different methods. The monochromatic μ_m and simulated polychromatic "apparent" μ_a values originate from thin beam spectrophotometer measurements, while the other two originate from the reconstructed images' circular ROIs. Note that the reconstructions closely follow the apparent μ_a of the solution when computed using a polychromatic model. The vertical error bars of the spectrophotometer-based estimations are smaller than the markers. The horizontal error bars are omitted for clarity.

and compare the outcome to the monochromatic attenuation as well as the reconstructed values. Another dataset used for the polychromatic absorption simulation was the camera's detector spectral response $W_c(\lambda)$. The relative response curve was provided by the manufacturer and proved quite steady (84%±7%) in the 550-700 nm wavelength interval. Considering the three available spectra, an estimate of F, the ratio of camera signals when exposed to the "attenuated" and the "unattenuated" light beams, was computed as follows :

$$F = \sum_{\lambda} \frac{W_e(\lambda) W_c(\lambda)}{\sum_{\lambda} W_e(\lambda) W_c(\lambda)} \exp\left(-\mu(\lambda) l_{\text{jar}}\right), \tag{3.6}$$

where l_{jar} is the jar diameter. The weights W were interpolated from the available data at 1 nm intervals. Based on the estimated signal fraction, an *apparent* μ, denoted μ_a was computed,

$$\mu_a = -\ln(F)/l_{\text{jar}}. \tag{3.7}$$

This is the value that a monochromatic algorithm such as FDK or OSC-TV is expected to yield from the DeskCAT scanner acquisitions. It was computed for all the dyed water jars and the values appear in Fig. 3.8(b). Note that this polychromatic estimate still neglects the effects of refraction, scatter, and nonlinearity of the camera response with respect to signal intensity.

The reconstructed μ_{FDK} and $\mu_{\text{OSC-TV}}$ estimation was carried on the central slice using a circular ROI of 5 cm in diameter. This ROI covered 55% of the jar cross section to exclude any

boundary artifacts. The estimate error for both algorithms was set to one standard deviation in the ROI. The results of the four types of μ estimation are summarized in Fig. 3.8(b). This figure leads to some important observations. FDK and OSC-TV perform similarly, with OSC-TV delivering marginally better results at lower dye volumes, and FDK doing so for higher dye volumes. The low accuracy of OSC-TV seems due to spectrum flattening. Since the FDK images are renormalized via least-squares, as explained in Sec. 3.5.5, they are equally subject to spectrum flattening. It is remarkable that generally, the reconstructed μ values are close to those estimated with the polychromatic model. This is consistent with the fact that refraction and stray light were minimal for this setup, which used a water tank and a water phantom (no gelatin added). The difference between the μ_m and μ_a ranges from 14% to 24%, which is a potential problem for the accuracy of any monochromatic reconstruction algorithm. A practical solution was proposed in the literature : an optical band-pass filter at the camera aperture could be used to narrow the signal spectrum for better conformity to the monochromatic model, see Olding *et al.* [40]. It should be noted that this approach reduces the projection data SNR and could undermine the quality of FDK images. Conversely, OSC-TV is expected to perform well despite noisy projection data. In this work, signal band-pass filtering was not implemented for the DeskCAT scanner. Hence, FDK reconstruction may potentially be modified by empirical projection data correction functions, which would require cumbersome calibration procedures. In contrast, a polychromatic iterative reconstruction algorithm could highly improve μ estimates based on a few spectrophotometer measurements.

3.6.4 3D mouse phantom analysis

The 3D mouse phantom projection data were reconstructed on a low-resolution grid of $192\times192\times64$ voxels using FDK and OSC-TV. The resulting axial and sagittal slices are shown in Fig. 3.9. This phantom is embedded in transparent gel and is itself composed of two gels of different μ values, one simulating soft tissue and the other simulating a highly attenuating contrast agent with a complex bowel-like shape. Small air bubbles had been left within the phantom during manufacturing, and caused significant refraction, appearing totally opaque on projections. As with the previous phantoms, rings and streaking are present in the FDK images. Rings are prominent in the axial slice and high-frequency streaking appears in the sagittal slice at the *bowel* level. The OSC-TV reconstruction algorithm effectively reduces such artifacts. An air bubble can be found at the lower-right corner of the sagittal slice. It created less perturbations in the OSC-TV image, similarly to the results obtained with the opaque line pair phantom. Nevertheless, thick high-density streaks connecting the bowels are visible in axial and sagittal slices and are equally present for both algorithms. A *black hole* is also observable in the axial slices, between the bowels. These artifacts are not effectively reduced by OSC-TV, indicating that opaque structures of large size remain challenging for this algorithm.

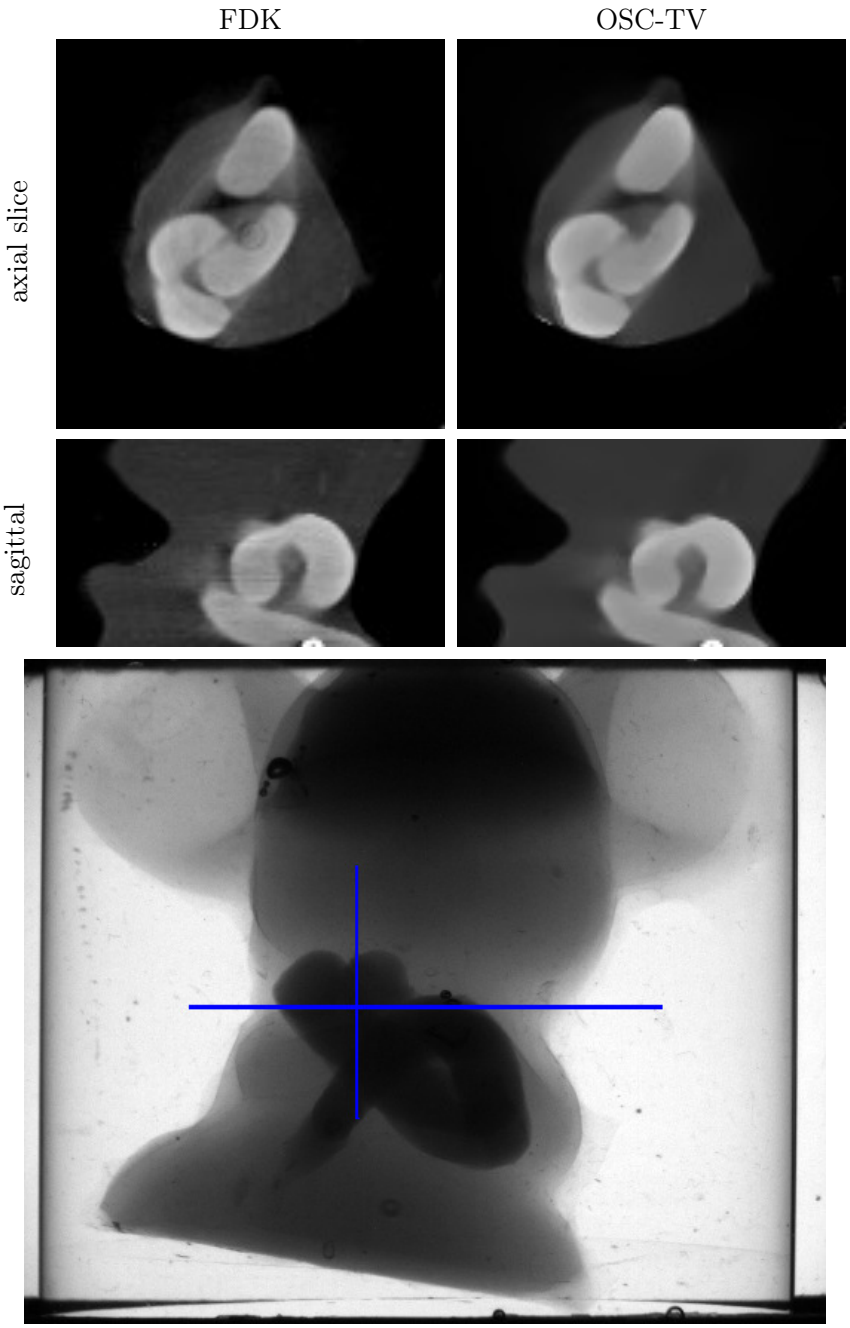


FIGURE 3.9 – Mouse phantom reconstructions and projection. First row : axial slice, taken 1.7 mm off-center, cropped to 120×120 voxels. Second row : sagittal slice, taken 7.1 mm off-center, cropped to 120×64 voxels. The μ interval is $[0.0, 1.75]$ cm^{-1}. Note the reduced ring and streaking artifacts when using OSC-TV. Bottom : DeskCAT camera snapshot of the phantom, raw intensity image. The horizontal line represents the location of the axial slice and the vertical line, the sagittal slice.

3.6.5 Reconstruction of small low-contrast structures

From a more general point of view, it should be noted that the design of the phantoms used in this study was relatively simple and their μ distributions were piecewise-constant. These objects are particularly well reconstructed by TV-regularized methods. Moreover, the phantoms did not contain objects of both small physical size and low relative contrast. To provide some insight into such cases, complementary tests were performed on a simulated human head phantom scan. The acquisition protocol is described in detail in section 2.5. Since the projection data was fully simulated, the modelling is identical for the optical and X ray imaging contexts. The tests revealed that physically small and low-contrast objects spatially resolved by the system kept their size, but lost relative contrast. For example, a line with a 2.6 mm cross section and a +4.3% contrast relative to the background was reconstructed using OSC-TV in a 256×256 mm^2 FOV. The reconstructed line spread to 2.9 mm (acceptable 11% increase), but its contrast fell more than two times to +1.8%. It suggests that OSC-TV is spatially edge-preserving, but small structures suffer from contrast loss. Other tests showed that for smaller objects with the same contrast, but below the spatial resolution limit, e.g., 1.5 mm width, the contrast became roughly averaged over a larger area of about 3 mm. The corresponding reconstructed images and profiles are provided in Appendix B. In summary, quantitative analysis of very small low-contrast structures is challenging for OSC-TV, but this algorithm enables the identification of such structures.

3.6.6 Reconstruction times

For the high-resolution grids, the OSC-TV reconstruction time was approximately 110 s using the GPU. The corresponding FDK executions on CPU required times of about 85 s. For the mouse phantom, the reconstruction times were of 93 s for OSC-TV and 32 s for FDK. Therefore, reconstruction times are within the same order of magnitude for both methods. Nevertheless, it is acknowledged that the FDK code was not fully optimized and used a single CPU core for execution. In general terms, both algorithms add negligible time to the gel reading procedure.

The scaling of the OSC-TV reconstruction time with respect to the number of detector rows, which influence the z-axis volume coverage for a given slice resolution, was also studied. For a high-resolution reconstruction grid of 320×320 voxel slices, the execution time was measured for three numbers of detector rows (30, 60, and 120) and was found to scale linearly in this interval, see Tab. 3.2.

The practical usefulness of the OSC-TV algorithm also depends on its implementation. The current CUDA code is relatively fast, with reconstruction times under two minutes for the largest volume grid of 320×320×76 voxels and a cone opening of 120 detector rows. Nevertheless, the precomputed and compressed projection matrix is stored in the GPU RAM memory,

TABLE 3.2 – Scaling of the OSC-TV reconstruction time for 320×320 slices, based on ten repetitions. The error was fixed to one standard deviation. The execution time scales linearly with the number of detector rows.

detector rows	volume slices	execution time (s)
30	20	28±1
60	38	55±2
120	76	111±2

limited to the 6 GiB of the Titan device. A 4× larger cone opening of 480 detector pixels would require 24 GiB GPU RAM, which is already available with the NVIDIA® Tesla® K80. The reconstruction time is estimated to be roughly double (about 4 min) instead of quadruple, because of the higher processing power of the K80 device. Alternative approaches require additional software development. A larger precomputed projection matrix can be partially cached in the GPU RAM memory and be regularly rewritten from the host personal computer (PC) RAM. This solution will increase the reconstruction time supralinearly because of additional memory operations. Another approach is to use the GPU raytracing to compute a very limited portion of the projection matrix, e.g., for one projection, and store it in the GPU RAM, eliminating transactions with the host PC RAM. Raytracing will require additional execution time, scaling linearly with the cone opening. For further acceleration, this approach may be extended to a multi-GPU computer, with the raytracing device computing the projection matrix and asynchronously feeding the result to the reconstructing device. In both cases, the reconstruction times are expected to scale linearly with the cone opening.

3.7 Conclusion

This paper presented a comparative evaluation of fully 3D tomographic reconstructions obtained via the OSC-TV and FDK algorithms in cone-beam optical CT. The results are mainly based on experimental projection data acquired on physical phantoms. Synthetic projection data and numerical simulations were also used to complement the discussion. The presented metrics quantitatively and qualitatively demonstrate that the OSC-TV algorithm outperforms FDK in cone-beam optical CT. This iterative approach overcomes a number of particular challenges of this imaging modality, improving spatial resolution and mitigating artifacts. Attenuation coefficients' estimation accuracy was found to be similar to that of FDK. Variable accuracy and uniformity of the reconstructed images were linked to beam polychromaticity. The image degradation caused by the jar wall effect seems persistent, although somewhat attenuated by OSC-TV. Considering that iterative reconstruction is progressively being integrated into x-ray CT systems, educational tools, such as the DeskCAT system, should reflect the specific properties of iterative algorithms. In this regard, GPU-accelerated OSC-TV yielded excellent illustrative results and could be employed in this field. The analysis of more

complex phantoms, like radiochromic gel dosimeters representative of radiation therapy dose distributions, is necessary to broaden the scope of the current results. The future work is also oriented toward accommodating larger datasets, as well as more accurate physics modeling and data precorrection. The implementation of a more sophisticated regularization method, better suited for 3D gel dosimetry, is also envisioned.

3.8 Acknowledgements

This work was supported in part by the Fonds de recherche du Québec – Nature et technologies (FRQ-NT). The authors acknowledge partial support by the CREATE Medical Physics Research Training Network grant of the Natural Sciences and Engineering Research Council of Canada (Grant No. 432290). They also wish to thank John Miller and Jen Dietrich from Modus Medical Devices, Inc. (London, ON, Canada) for their extensive advice regarding the DeskCAT system specifications, as well as Marie-Ève Delage and Patricia Duguay-Drouin for their help with the DeskCAT emission spectrum measurement.

Chapitre 4

System matrix computation vs storage on GPU : a comparative study in cone beam CT

Dmitri Matenine[1], Geoffroi Côté[1], Julia Mascolo-Fortin[1], Yves Goussard[2] et Philippe Després[1,3,4]

[1] *Département de physique, de génie physique et d'optique, Université Laval, Québec (Québec), Canada*

[2] *Département de génie électrique / Institut de génie biomédical, École Polytechnique de Montréal, Montréal (Québec), Canada*

[3] *Centre de recherche sur le cancer, Université Laval, Québec (Québec), Canada*

[4] *Département de radio-oncologie et Centre de recherche du CHU de Québec (Québec) Canada.*

4.1 Résumé

Cet article présente une comparaison de méthodes de stockage et de calcul de la matrice-système en reconstruction itérative sur GPU pour la TDM à faisceau conique. Compte tenu du besoin de la reconstruction itérative d'effectuer des projections radiographiques simulées et des rétroprojections multiples, la performance numérique de ces opérateurs mathématiques est cruciale pour assurer un temps de reconstruction raisonnable. Plus précisément, dans ce travail, la matrice-système modélise l'intersection de rayons fins et des voxels sous forme de blocs placés sur une grille cartésienne, une représentation relativement fidèle de la géométrie d'acquisition. Toutefois, une telle matrice, sans compression, dépasse la taille de la mémoire vive des ordinateurs typiques par un facteur dix ou plus. Compte tenu des limites de la mémoire vive du GPU, plusieurs méthodes de traitement de la matrice système ont été comparés : le stockage complet d'une matrice-système compressée, le calcul de ses éléments à la volée, ainsi

que le stockage partiel de la matrice-système combiné au calcul à la volée au besoin. Ces approches ont été testées sur des géométries représentant une acquisition TDM à faisceau conique d'une tête humaine. Les temps d'exécution de trois opérateurs ont été comparés : la projection directe, la rétroprojection et l'itération de l'algorithme itératif OSC. La méthode de stockage complet a donné lieu aux temps d'exécution les plus courts de la rétroprojection et de l'itération OSC, avec une accélération de 1.52× par rapport au calcul à la volée pour OSC. Toutefois, cette approche est limitée par la mémoire vive GPU disponible et les symétries géométriques qui assurent la compression de données. Le calcul a la volée a donné lieu aux projections directes les plus rapides, ainsi que des temps de calcul raisonnables pour l'itération OSC, avec 176.4 ms par angle d'acquisition et par itération, pour le plus grand détecteur testé, sans recours à quelque symétrie que ce soit. Le stockage partiel de la matrice système a montré une performance en temps semblable à celle de la méthode à la volée, tout en conservant la dépendance aux symétries, se révélant comme la méthode la moins intéressante. En somme, la méthode de traçage à la volée s'est montrée la plus flexible, alliant des temps de calcul raisonnables et l'indépendance par rapport aux symétries. La méthode avec stockage complet, étant la plus rapide, pourrait être d'intérêt pour certaines applications sur mesure axées sur la performance numérique.

4.2 Abstract

Purpose : Iterative reconstruction algorithms in computed tomography (CT) require a fast method for computing the intersection distances between the trajectories of photons and the object, also called ray-tracing or system matrix computation. This work focused on the thin-ray model and is aimed at comparing different system matrix handling strategies using graphical processing units (GPUs).

Methods : In this work, the system matrix is modeled by thin rays intersecting a regular grid of box-shaped voxels, known to be an accurate representation of the forward projection operator in CT. However, an uncompressed system matrix exceeds the random access memory (RAM) capacities of typical computers by one order of magnitude or more. Considering the RAM limitations of GPU hardware, several system matrix handling methods were compared : full storage of a compressed system matrix, on-the-fly computation of its coefficients, and partial storage of the system matrix with partial on-the-fly computation. These methods were tested on geometries mimicking a cone beam CT (CBCT) acquisition of a human head. Execution times of three routines of interest were compared : forward projection, backprojection and ordered-subsets convex (OSC) iteration.

Results : A fully stored system matrix yielded the shortest backprojection and OSC iteration times, with a 1.52× acceleration for OSC when compared to the on-the-fly approach. Nevertheless, the maximum problem size was bound by the available GPU RAM and geometrical

symmetries. On-the-fly coefficient computation was shown to be the fastest for forward projection. It also offered reasonable execution times of about 176.4 ms per view per OSC iteration for the largest dataset tested, without requiring symmetries. Partial system matrix storage has shown a performance similar to the on-the-fly approach, while still relying on symmetries.

Conclusion : Partial system matrix storage was shown to yield the lowest relative performance. On-the-fly ray-tracing was shown to be the most flexible method, yielding reasonable execution times. A fully stored system matrix allowed for the lowest backprojection and OSC iteration times and may be of interest for certain performance-oriented applications.

keywords : cone-beam CT, ray-tracing, iterative reconstruction, graphics processing units

4.3 Introduction

Computer tomography (CT) is a three-dimensional (3D) imaging modality with a wide range of applications. It is based on measurements of the partial attenuation of a beam of radiation in the imaged subject. X-ray CT was first developed for medical imaging, then similar technologies emerged for non-destructive material testing and specialized application areas.

The development of practical designs of CT systems closely followed the introduction of digital electronic computers. Reconstruction methods necessary for CT imaging are roughly classified as analytical and iterative. Iterative reconstruction (IR) algorithms often rely on a refined physical and geometric model (system matrix), while analytical algorithms mostly use interpolation. Also, analytical algorithms perform a single backprojection of the acquired projection data, while IR methods gradually refine the 3D image. IR was shown to outperform analytical methods in many aspects and is highly desirable in clinical applications [6, 18], but its practical implementation is often hindered by numerical complexity.

When designing an iterative algorithm, one should select an appropriate mathematical representation of the system matrix. IR expects an accurate geometrical representation to ensure convergence [143], thus generally leading to more geometry-related computations. The thin-ray rectilinear radiation propagation model was shown to be an accurate system matrix representation for forward projection [73, 174], and is therefore used in the current work. It is also characterized by high computational complexity, compared to voxel-driven interpolation methods [4] and advanced approximate weighting methods, such as the distance-driven [77] and the separable footprints [78] algorithms.

The Siddon's algorithm [79], in its *incremental* version [80, 81] implemented on GPU, was used to compute the intersection distances, alternatively called *weights*. In this work, one line segment is traced between the radiation source and the center of each detector element. A similar and more sophisticated model, sometimes called multi-ray-tracing, uses several segments per detector element and provides better model accuracy at the cost of longer computation

times [73]. The proposed technique is trivially extensible to this formalism, but the estimated computation times still appear prohibitive for multi-ray-tracing.

For current high-resolution clinical applications, the system matrix is difficult to store in the random access memory (RAM) of a computer in its entirety [87,88]. Memory bandwidth may also be a concern when relying on repetitive data retrieval. To avoid system matrix storage altogether, advanced interpolation algorithms were designed to significantly reduce on-the-fly computation time of the weights [77,78]. This work is focused on the computation of thin-ray intersection distances, with a relatively accurate forward-projection modelling and reasonable ray-tracing time in mind.

Repetitive projection and backprojection routines often represent the numerical bottleneck of IR due to the large datasets involved and complex memory access patterns imposed by the acquisition geometry. It should be added that most model-based IR algorithms are not *embarrassingly parallel* i.e., their parallelization requires an important design effort, as well as architecture-aware optimization. For example, certain algorithmic accelerations suitable to CPU computing, like ordered subsets, may impede the massive GPU parallelism and should be fine-tuned for GPU, as shown by Xu *et al.* [175]. The influence of numerical precision on IR was investigated by Maaß *et al.*, indicating that low-precision data representation may reduce reconstruction time without significantly affecting image quality [102]. Furthermore, IR implementations using multi-GPU hardware were recently reviewed by Jia *et al.*, revealing the inherent complexity of inter-GPU communication and synchronization in the tomographic reconstruction context [24]. Considering the correlation between absolute performance and development complexity, this comparative work is rather focused on evaluation of ray-tracing using a single GPU.

4.4 Theory

4.4.1 Basic problem

Iterative reconstruction is often based on a relatively accurate physical model of the radiation attenuation process, which is used during simulated projection of the radiation through the 3D image estimate. Particles follow rectilinear trajectories from the radiation source, intersect the subject represented by a discrete grid of attenuation values, and reach bins of a discrete two-dimensional (2D) detector grid, thus forming X ray *projections* of the object. Using several incidence angles, several projections are acquired and form a *sinogram*, which is vectorized into a column vector $\mathbf{y}$, where i denotes the reading index and is also a unique identifier to the corresponding ray trajectory. Similarly, the image grid is expressed as a column vector μ, indexed by j, the unique voxel index. The problem of interest in this paper is to generate a system matrix $\mathbf{A}$ of ray-voxel intersections, with its elements l_{ij} being intersection distances. $\mathbf{A}$ is essential to simulate the imaging process, which may be denoted as a matrix-vector

product :

$$\mathbf{y} = \mathbf{A}\mu. \tag{4.1}$$

The formulation of backprojection equations depends on the type of estimator and the optimization method, but also involves the l_{ij} coefficients to weigh the correction terms of the values of image voxels.

4.4.2 System matrix models

For a majority of numerical models of the imaging process, voxels are box-shaped and are assumed to possess a uniform linear attenuation coefficient μ. Box-shaped voxels are convenient from a mathematical and numerical standpoints due to their equal size and simple indexing. However, the representation suffers from partial volume artifacts for high-contrast interfaces [9]. Alternative representations of the volume grid include lattices based on the cylindrical coordinates [67–69]. The goal pursued by this representation is extreme system matrix compression via rotational symmetry ; nonetheless, this representation influences IR convergence properties and is still under investigation. Another image model is the superposition of spherically symmetric basis functions [70–72], which have analytically described footprints on the CT detector. Such methods are known for reducing aliasing artifacts thanks to continuous object modelling.

With respect to radiation, rectilinear trajectories do not properly model X-ray scatter. However, most software-based scatter compensation methods rather correct the detector readings, regardless of radiation paths used for backprojection, leading to a satisfactory balance of accuracy and complexity [114].

Geometrical hypotheses determine the resulting weights and the numerical complexity of the model. The basic approach in CT is voxel-driven bi-linear interpolation and assigns relative weights based on the position of the center of a voxel and the intersection of a ray-line and four neighbouring detector pixels. Another well-known interpolation-based formulation is the Joseph method [76]. The thin-ray model computes the euclidean distance between the entry and exit points of an infinitely thin ray in a voxel [79]. Another approach is to find a 3D intersection volume between the voxel and the pyramid formed by the source and a rectangular detector bin, known as the box-beam method [73, 82, 83, 176, 177]. The thin-ray model was shown to be the most accurate when compared to subpixel-based linear interpolation methods in terms of projection and backprojection accuracy, by Zhuang *et al.* [174]. Xu and Mueller compared several interpolation and integration methods applied to forward projection and simultaneous algebraic reconstruction technique (SART) [73]. The relative accuracy of the methods was found to depend on the model used to generate synthetic raw data for their simulation study ; therefore, the thin-ray model was significantly superior to interpolation in some cases, but rather equivalent in other cases.

Advanced interpolation methods proposed in the recent years offered improved balance between accuracy and computational performance, by providing simple weighting functions computed on-the-fly, avoiding system matrix storage. The distance-driven (DD) method proposed by De Man and Basu [77] and the separable footprints (SF) method by Long *et al.* [78] are well established in CT. These methods are significantly more accurate than bi-linear interpolation, as shown by the comparative evaluation by Karimi and Ward [143], with SF being more accurate than both DD [78, 143] and footprints of spherically symmetric basis functions [143]. Nevertheless, the relative error introduced by both methods varies with the projection angle for typical CBCT geometries and the weights are again relative. Another study by Hahn *et al.* [144] compared DD, the Joseph method and bi-linear interpolation. It demonstrated that DD was superior to the Joseph method, at a greater computational cost, and a choice between the two may be based on the acquisition parameters. Thin-ray tracing, for its part, is very accurate for forward projection, but might lead to artifacts when weights are reused for backprojection [178]. This drawback is usually countered via image regularization and multi-ray-tracing. Multi-ray-tracing is sometimes used as a gold standard in terms of accuracy of the forward projection model [73], to approximate a continuous acquisition model using oversampling [159], so a high-performance implementation of such operators may benefit practical applications of IR.

4.4.3 Thin-ray tracing algorithms

The basic approach to find the intersection distances defined by the thin-ray model is exposed in Siddon's ray-tracing algorithm [79]. This algorithm was improved in different ways, leading to the *incremental* formulation [80, 81]. The latter was evaluated in terms of numerical performance using central processing unit (CPU) -based hardware [85], and GPU implementations of the method were also proposed [86, 179]. A set of promising alternative projection/backprojection algorithms were proposed by Gao [146]. They are characterized by a theoretical complexity of $\mathcal{O}(1)$, compared to $\mathcal{O}(n)$ for Siddon's tracing, where n denotes the number of voxels along an edge of a cubic image grid. $\mathcal{O}(1)$ complexity is achieved by parallelizing the tracing loop over a single full ray. Experimentally, this algorithm was shown to provide excellent occupancy for typical GPUs even without parallelizing the loops over full rays. It was demonstrated to outperform the basic Siddon's algorithm in terms of execution time ; nevertheless, it was not compared to its incremental version, which is inherently faster. Another GPU-accelerated thin-ray matched projector/backprojector pair was proposed by Nguyen and Lee [147]. It uses a geometric formalism to efficiently confine the footprint of a given voxel on the detector grid and thus enable voxel-driven backprojection. It was shown to offer reasonable execution times, being about 4 times slower than a typical mismatched operator pair (a thin-ray projector and interpolation-based backprojector).

To generate system matrix data, the common and relatively efficient incremental Siddon's

method was used here. Formally comparing the above-mentioned implementations of the thin-ray model is out of the scope of this paper, since it is rather focused on the effects of system matrix storage on the projection and backprojection operators. Notwithstanding, the presented work was put into the perspective of the recent findings by Nguyen and Lee [147] in Sec. 4.6.3.

4.4.4 System matrix storage

For helical CT, a few studies investigated a stored-matrix approach. The numerical handling of the thin-ray system matrix was studied by Xu and Tsui [87] using CPU-based computing. Their work relied on geometrical symmetries specific to helical CT to achieve reasonable system matrix size. Execution times ranged from 7 min to 6 h per projection-backprojection pair, depending on problem size. System matrix storage using a formalism of shift invariance in rotating coordinates was proposed by Guo and Gao [180], achieving acceleration factors of 3 to 6 for thin-ray projection and 3 to 16 for interpolation-based backprojection, compared to on-the-fly implementation of the same operators.

For CBCT, sparse matrix compression applied to iterative reconstruction was investigated by Jian-lin *et al.* [88], applying different storage orders to projection and backprojection matrices, thus achieving data compression factors ranging from 128 to 275. The proposed study is focused on a numerical comparison of a few system matrix handling methods in CBCT which could be readily integrated in a variety of IR algorithms, namely full system matrix storage, partial system matrix storage and on-the-fly ray-tracing.

4.5 Materials and Methods

4.5.1 Special nomenclature

This paper is focused on ray-driven projection and backprojection, and a notation convenient for this formalism is proposed below. For a rectilinear ray crossing several voxels, let k denote the sequential index of a voxel crossed by a ray, not to be confused with j – a unique location in the volume. For each attenuated ray i, a vector of indices $\mathbf{j}_i$ of the crossed voxels and a vector of intersection distances $\mathbf{l}_i$ is necessary for projection and backprojection. Let a *system sub-matrix* designate a set which contains the above data for a single incidence angle ϕ :

$$\mathbf{M} \equiv \{\{\mathbf{j}_i, \mathbf{l}_i\} | \phi = \text{const}\}. \tag{4.2}$$

A sub-matrix is typically one to two orders of magnitude lower in size than the entire system matrix and may be stored in GPU RAM.

4.5.2 Overview of the system matrix handling methods

The fully stored and compressed system matrix method considered here was described in detail by Matenine et al. [181]. It consists of pre-computing the system matrices for a number of

typical acquisition protocols and storing them on disk. The compression scheme stores data in three vectors. The first vector contains the number of voxels crossed by each ray of the acquisition and is only needed for proper data access in the two larger vectors. The second vector contains non-zero l_{ij} values i.e., a continuous sequence of $\mathbf{l}_i$ vectors. A third vector stores the respective indices of the crossed voxels necessary for projection and backprojection i.e., a continuous sequence of $\mathbf{j}_i$ vectors.

The on-the-fly (OTF) system matrix generation, in this study, consists of performing the incremental Siddon algorithm loop for each ray i, where the reference j and length l_{ij} are local variables inside the loop (not data vectors) and therefore cannot be reused. In consequence, the tracing loop is performed for projection and again for backprojection.

The partial storage of the system matrix may be interesting when the GPU RAM is insufficient to fully store $\mathbf{A}$. It consists of filling all the available GPU RAM with the maximum possible number of system sub-matrices $\mathbf{M}$ before reconstruction. Since IR implies several projections and backprojections of the entire sinogram, the same sub-matrices are stored for the entire reconstruction and are guaranteed to be reused at each iteration. For views without a stored sub-matrix, the method falls back on OTF ray-tracing.

4.5.3 CUDA architecture

The parallel code was developed using NVIDIA® (Santa Clara, California) GPUs and the CUDA™ programming interface. CUDA is a single-instruction-multiple-data (SIMD) architecture, which consists of launching several thousands of computing *threads* simultaneously. Threads are grouped in *thread blocks* of several hundred threads, which execute on streaming multiprocessors (SMs), while several blocks constitute a *grid*. A *kernel* is a unique code which will be executed by each thread of the grid. Each thread block has access to on-chip shared memory, with sizes in the tens of kB per SM. Large datasets are stored in the GPU RAM, usually called *global memory*, with sizes of 1-12 GiB per GPU.

4.5.4 Time metrics

In this paper, the relative performance of the following computational tasks were evaluated : forward projection, backprojection and OSC iteration. The first two evaluate general-purpose tasks used in IR, while the third serves as an example of a complete reconstruction algorithm. The implementation of OSC required a projection followed by a backprojection into two accumulating matrices, since there are two weighted sums in the voxel update equation. The numerical results of each task were written to GPU RAM and this operation was accounted for in the time measurement. For the OSC iteration, the number of subsets was irrelevant here, since only the time-consuming backprojections to accumulating matrices were benchmarked, while the μ updates were of negligible cost.

4.5.5 Parallelization

The fully pre-computed system matrix method is parallelized as follows : a CUDA thread block is assigned with one or few rays to process. Each thread is assigned with one voxel of the ray or a few voxels of the same sequential index k, see Fig. 4.1 (a). The corresponding sets $\{\mathbf{j}, \mathbf{l}\}_i$ are loaded into shared memory. The possibility of assigning several rays per block is intended to maximally fill the faster shared memory. Forward projection, which is a dot product of $\mathbf{l}_i$ and the corresponding μ values, is performed by the threads of the block via parallel reduction, a divide-and-conquer technique [95]. Backprojection is performed in parallel for all the concerned voxels. Race conditions at backprojection are avoided by running the kernel over a group of rays of a single view which are guaranteed to write to separate locations in μ [181]. In the case of the OSC iteration, a single kernel executes the projection and backprojection, thus reusing the $\{\mathbf{j}, \mathbf{l}\}_i$ stored in shared memory. This strategy is referred to as *block-based*.

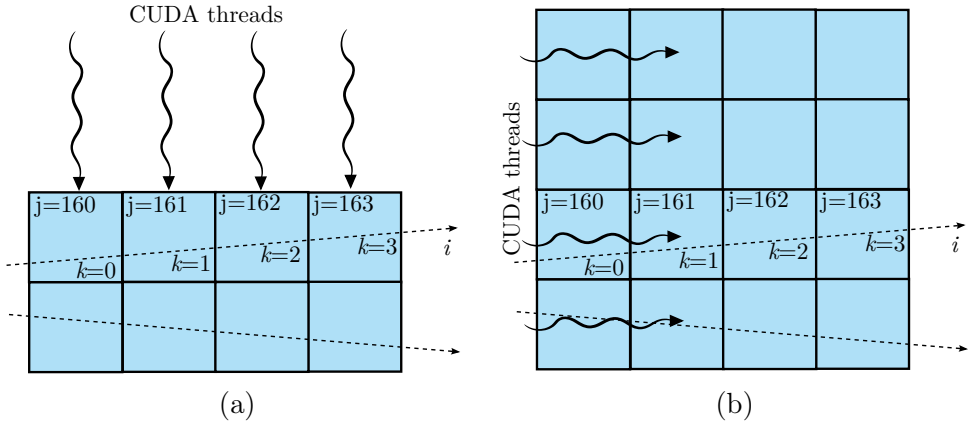


FIGURE 4.1 – Parallelization strategies. (a) *Block-based* method for the fully stored and partially stored system matrix : a thread block acts on a few rays and each thread acts on a few voxels of index k. (b) *Loop-based* method for the on-the-fly operators : each thread acts on all the voxels of a ray i using a loop. Notation : i is a unique ray index, j is a unique voxel index and k is the sequential index of a voxel crossed along a ray.

When using OTF system matrix generation, the projection and backprojection routines are parallelized using one thread per ray. Each thread executes a loop to compute the distances and references while projecting or backprojecting, see Fig. 4.1 (b). In the case of the OSC iteration, the tracing loop is called once for projection and a second time for backprojection. This strategy is referred to as *loop-based*.

For the partial system matrix storage, two parallelization strategies were implemented and compared : a thread block for a few rays similar to the fully stored system matrix parallelization and one thread per ray with a loop over the ray, similar to the OTF parallelization.

4.5.6 Memory coalescence

The CUDA hardware architecture reads/writes data from/to GPU RAM by simultaneously transferring a few hundred bytes per memory transaction. Different threads with contiguous ID indices should access contiguous memory elements to ensure maximum bandwidth ; this concept is called *memory coalescence* [95]. In this work, it was taken in consideration regarding access patterns of the system matrix and the reconstructed 3D volume.

With respect to system matrix, for cases involving one thread block per few rays, coalescence was ensured by storing $\{\mathbf{j}, \mathbf{l}\}_i$ linearly, since they are accessed by contiguous threads working on the same ray. For the case of partial system matrix storage using one thread per ray, the same order was empirically found to offer the lowest execution times.

With respect to the reconstructed 3D volume, coalescence was ensured via the reordering of the volume into *subvolumes*, as described in detail by Matenine *et al.* [181]. A typical subvolume of a size of $4 \times 4 \times 2$ voxels is stored in linear memory, so contiguous threads update contiguous voxels. This volume reordering is expected to be efficient for all the methods above. For block-based methods, contiguous threads write to contiguous voxels *along* the ray, while for the loop-based methods, the contiguous threads write to adjacent voxels *across* adjacent rays.

4.5.7 Geometric symmetries

Several symmetries may be exploited when computing and partially storing the system sub-matrices as proposed above. A basic usage of symmetries was presented by Matenine *et al.* [181] and a slightly extended formalism is presented in this paper. A more detailed description of the conditions which ensure different symmetries may be found in Appendix C. The most likely one is the z-axis mirror symmetry, and requires that the source-detector axis reaches the central detector row. This reduces the size of the sub-matrix by a factor of two. The next symmetries require a square slice profile centered on the rotation axis, as well as isometric voxels in-slice. With an integer number of regularly spaced acquisition angles covering exactly 90°, the rotational symmetry reduces the total number of *unique* sub-matrices by a typical factor of 4. If the source-detector axis reaches the center of a detector row, the xy plane symmetry may be used to reduce the number of sub-matrices to those that cover the first 45° of the acquisition. In consequence, a stored unique sub-matrix may be valid for up to 8 projection views. Combined with the z-axis symmetry, system matrix data may be reused up to 16× per iteration. It is important to note that the symmetries are not necessary for OTF ray-tracing, since the j and l_{ij} are stored as local variables for each ray trajectory. In contrast, symmetries are crucial for stored-matrix methods, since the pre-computed data may be reused. Hence, in this study, all of the symmetries were valid for the benchmarks involving fully or partially stored system matrices.

4.5.8 Projection data and image grid

The simulated geometry was similar to the Varian® (Palo Alto, California) OBI CBCT system, in particular, the head scan setup. For simulated projections, the source-isocenter and the detector-isocenter distances were set to 100 cm and 56 cm respectively. The detector was simulated as a flat panel with square pixels, with a detector row length of 39.7 cm in the *fan* direction, corresponding to 512 pixels. The problem size was varied via the *cone opening*, measured in pixels. A first set of cone openings was selected so as to be compatible with the fully pre-computed system matrix and the total GPU RAM : 32, 64, 96, 128, 160 and 192 pixels. This set was also useful for comparisons on lower-end hardware. In order to evaluate more realistic use cases, wider cone openings of 224, 256, 288, 320, 352, 384, 416 and 448 pixels were added to the former set. The reconstructed slice grid was square, with a side length of 25.6 cm and the corresponding image grid was of 384×384 cubic voxels with 0.667 mm sides. The volume z-span was computed automatically in accordance with the cone opening, to avoid major truncation artifacts, and varied from 28 to 384 voxels. The projection dataset consisted of 200 projections acquired at regular intervals, which represents a few-view acquisition. For comparison, a standard-dose head acquisition with the OBI scanner features about 360 projections, while a standard abdomen-pelvis scan features about 700 projections.

The reported execution times are the average of 5 runs. The projection, projection-backprojection and OSC operators were run for the 200 projections at their respective incidence angles, so an average execution time per projection could be computed. Total GPU memory usage was registered for all the methods. For partial storage methods, the system matrix storage coefficient i.e., the fraction of system matrix pre-computed and stored in GPU RAM was evaluated.

4.5.9 Computing hardware

Two NVIDIA® (Santa Clara, CA) GPUs were used to perform the benchmarks : A GeForce® GTX 580 and a GeForce® Titan™, with their main characteristics shown in Tab. 4.1. The

TABLE 4.1 – GPU hardware specifications. Raw computing power is expressed in giga- floating-point operations per second (GFLOPS), for 32-bit arithmetic in this case.

Card	GTX 580	Titan
Cores	512	2688
GFLOPS	1581	4500
RAM (GiB)	1.5	6.0
Bandwidth (GiB/s)	192	288

host computer was equipped with an Intel® Core™ i7-960 CPU with a 3.20 GHz clock and 12 GiB RAM. 32-bit precision was used for all floating-point computations.

4.6 Results and Discussion

4.6.1 Execution time

The progression of execution time was plotted with respect to the problem size for the different methods. The results for the GTX 580 are shown in Fig. 4.2. Due to the limited GPU RAM, the fully pre-computed system matrix method was not tested on this GPU. For the tested methods, the maximum cone opening was of 192 pixels. OTF ray-tracing performed best for forward projection with 11.4 ms per view for the maximum cone opening, while partial storage with multiple threads per ray was by far the slowest approach (18.4 ms). The partial storage method with the loop-based kernel tended towards the OTF curve. For backprojection, partial storage with the loop kernel performed marginally better than the OTF kernel, for larger cone openings. The backprojection time was of about 82 ms for these methods for the largest cone opening. The block-based kernel was again the slowest. The relative performance of the three methods was similar for OSC, yielding about 150 ms per projection for the maximum cone opening.

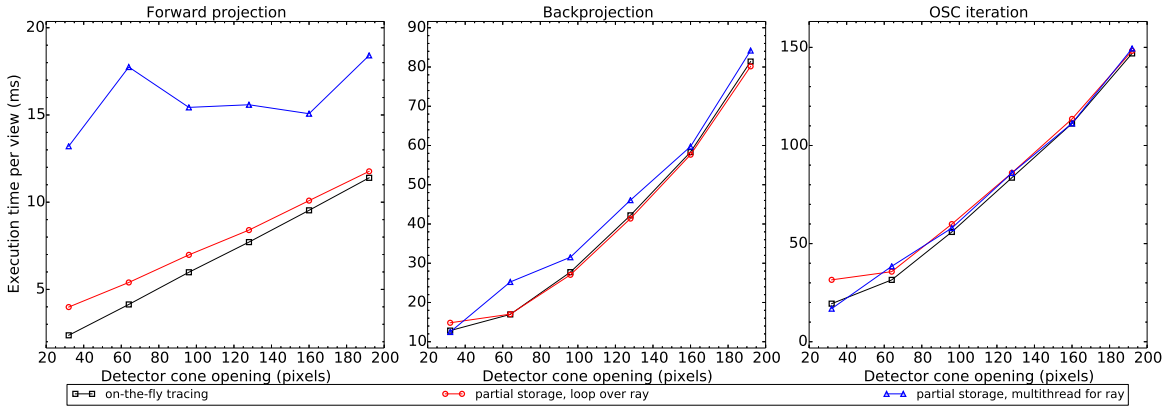


FIGURE 4.2 – Time metrics for the GTX 580 GPU. On-the-fly ray-tracing performs best for forward projection. Partial storage with the loop kernel is marginally better for backprojection. Comparative performance of the methods is similar for OSC. The error due to variability between runs is smaller than the marker size.

For the Titan GPU, execution times are presented in Fig. 4.3. For forward projection, OTF ray-tracing was again the fastest approach, with 14.7 ms per view for a cone opening of 448 pixels. It was closely followed by the partial storage method using the loop kernel, with 16.1 ms. The fully pre-computed system matrix method yielded 26.2 ms for a cone opening of 192 pixels, while the partial storage method with block-based kernel was the slowest, with 38.8 ms for a cone opening of 448 pixels. For backprojection, the fully pre-computed system matrix method was the fastest with 26.4 ms per view, although the GPU RAM limited the treatable problem size to a cone opening of 192 pixels. The partial storage method using the

loop-based kernel performed marginally better than the OTF ray-tracing and both yielded about 90 ms for a cone opening of 448 pixels. Partial storage with the block-based kernel was by far the slowest approach for backprojection, yielding 103.8 ms. For the OSC iteration, the fully pre-computed system matrix approach was again the fastest, yielding 44.2 ms for its maximal cone opening of 192 pixels, vs. 67.2 ms for OTF. OTF ray-tracing was the fastest approach for larger problems, with 176.4 ms per view at 448 pixels, while partial storage methods yielded the longest reconstruction times of about 190 ms per view.

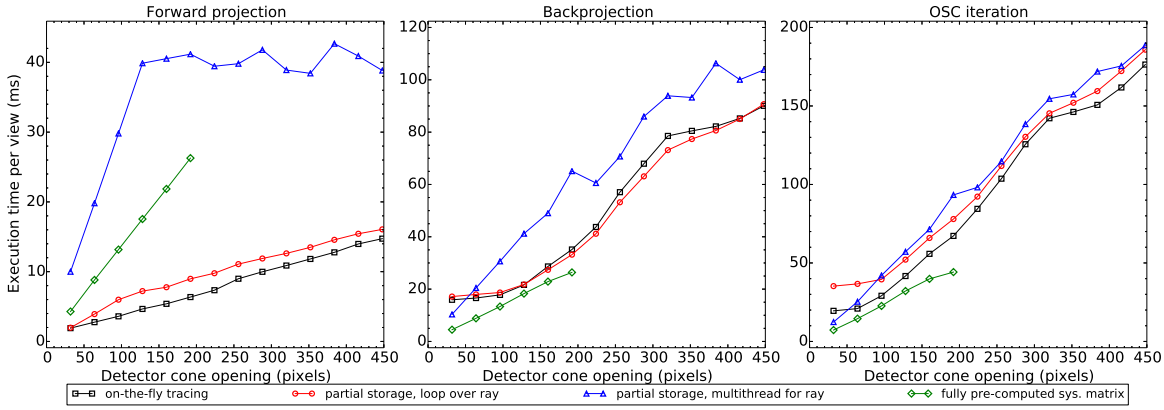


FIGURE 4.3 – Time metrics for routines over one projection view for the Titan GPU. On-the-fly ray-tracing performs best for forward projection. Fully pre-computed system matrix performs best for backprojection and OSC, but maximal cone opening is limited. Partial storage with loop kernel performs similarly to the on-the-fly method and the block-based kernel is slowest. The error due to variability between runs is smaller than the marker size.

The variability between runs (standard deviation for 5 runs) was of the order of 0.01 ms or lower and was not plotted. The execution time variability between different incidence angles yielded a standard deviation of about 20% ; however, this variability did not affect execution time for a whole sinogram and was not plotted either.

Comparing forward- and back- projection operators, backprojection is slower for all methods due to a ray-driven formulation of the problem, which limits practical memory coalescence and constitutes the major bottleneck. Forward projection using the fully stored system matrix is slower than OTF, likely due to the overhead needed to retrieve system matrix data. Fast OTF forward projection also points to the fact that $\{\mathbf{j}, \mathbf{l}\}_i$ computation is not costly *per se*, especially relative to the complete backprojection operator. In consequence, it appears that the combination of a fully pre-computed system matrix and a block-based parallelization leads to a maximum bandwidth at backprojection, while the loop-based parallelization lags in this regard. More generally speaking, regardless of implementation, the reconstruction problem is rather bandwidth-bound.

4.6.2 Memory requirements

Another metric to consider is the total memory requirement of the different methods. All the data structures necessary for the OSC reconstruction algorithm were allocated to measure the GPU RAM requirements, see Fig. 4.4, to produce the most realistic assessment. For the

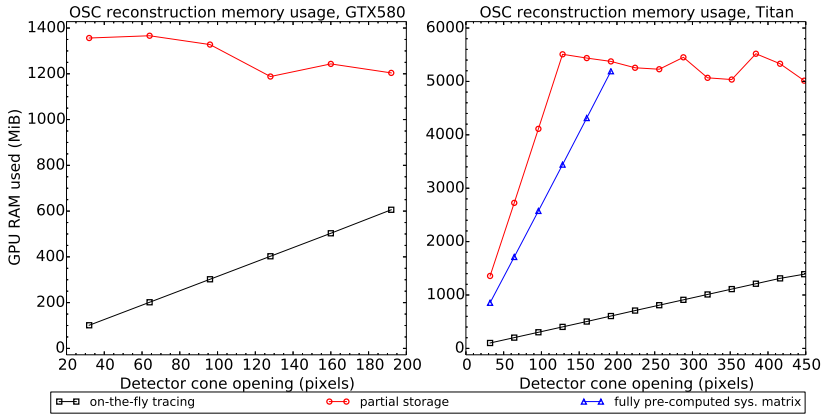


FIGURE 4.4 – Total GPU RAM memory usage for OSC reconstruction. Note that the pre-computed method tested on the Titan GPU is not applicable beyond a cone opening of 192 pixels. On-the-fly raytracing requires the least memory resources.

pre-computed method (on Titan GPU only), the memory usage increased with a high constant slope, so that the method became unusable beyond a cone opening of 192 pixels. For OTF ray-tracing, memory usage was low, at most 1.4 GiB for a cone opening of 448 pixels. For partial storage methods, almost all of the GPU RAM was purposely filled with system matrix data ; therefore, memory usage was high and practically constant for most of the cone openings and both GPUs tested. The memory usage was higher for partial storage methods than for the fully pre-computed method ; this is due to lower compression levels of the system matrix which simplify array indexing. An additional metric important for the partial matrix storage methods is the system matrix storage coefficient, see Fig. 4.5. For the GTX 580 GPU, it quickly dropped from 1.0 to 0.077 as the problem size grew. For the Titan GPU, a similar drop from 1.0 to 0.192 was observed between cone openings of 128 and 448 pixels. In practical terms, a low storage coefficient means that the partial storage method fell back on the OTF kernel for the majority of the views. This explains the high performance of the forward projection using partial storage and a loop-based kernel.

4.6.3 General discussion

In terms of relative performance, the fully pre-computed system matrix approach has shown the shortest OSC iteration times for the treatable problem sizes e.g., a 1.52× acceleration compared to the OTF kernel for a cone opening of 192 pixels. This speed advantage of a pre-computed system matrix for CBCT reconstruction confirms recent results by Guo and

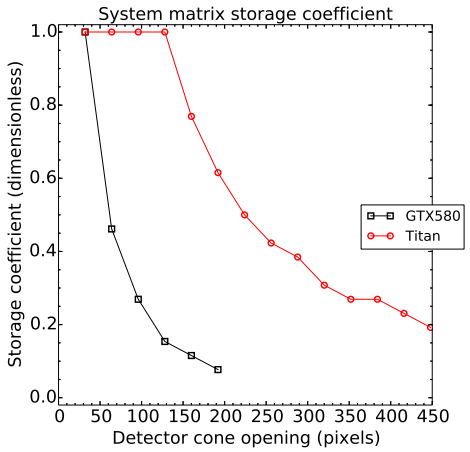


FIGURE 4.5 – System matrix storage coefficient for partial storage methods. It is trivially 1 for the fully pre-computed method and 0 for on-the-fly ray-tracing.

Gao [180] for helical CT geometry. However, GPU RAM was shown to be the major limitation of our pre-computed approach, as well as reliance on several geometrical symmetries which may not be available with certain CBCT devices and protocols. On-the-fly ray-tracing approach was shown to offer reasonable relative performance, while being the most flexible, using very low GPU RAM and independent of any symmetry properties. It was also shown to be the fastest for forward projection, therefore, a method of choice for mismatched projector-backprojector pairs in iterative CT reconstruction. The partial storage methods were shown to be a poor implementation choice, since they inherit the constraints of the fully pre-computed method without any significant performance gain, even when compared to the OTF approach. In sum, it appears that the OTF approach is an appropriate choice for a general-purpose reconstruction code or a CT reconstruction routine library, while the fully pre-computed system matrix could be employed to provide an additional performance gain for niche systems which satisfy the constraints imposed by this method.

In absolute terms, the execution times reported in this paper are of the same order of magnitude as those presented in recent studies. For the OSC step, the OTF kernel yielded 176.4 ms per view for 512×448 detector and a 384^3 image grid, using a GeForce Titan GPU. For comparison, the thin-ray projector/backprojector by Nguyen and Lee [147] yielded OSC iteration times of about 1300 ms per view for a 784×964 detector and a 512^3 image grid, using a GeForce GTX 680 GPU. In this case, the number of detector pixels is about 3.3× larger, and the GPU GFLOPS metric and memory bandwidth are each about 1.5× lower, which explains the longer execution time.

The results of this study seem independent of the hardware : the OTF kernel generally performs better than partial storage kernels for both GPUs tested. The apparent performance gap is

narrower for the GTX 580 card, which is due to low system matrix storage coefficients : since the partial storage code called mainly OTF kernels, it yielded performance similar to OTF.

4.7 Conclusion

This paper compared different system matrix handling methods for iterative reconstruction in cone beam CT using GPU, using a geometric model combining thin rays and box-shaped voxels. Three major approaches were compared : fully pre-computed and stored system matrix, on-the-fly (OTF) ray-tracing, and partial storage of system matrix with fallback on OTF ray-tracing. The fully pre-computed method was shown to be 1.52× faster than OTF for the ordered subsets convex (OSC) algorithm, although the maximum problem size was constrained by total GPU memory and geometrical symmetries. The acceleration was due to a better bandwidth usage for a bandwidth-bound problem. The OTF approach was shown to be the most flexible, using low GPU memory and being independent of symmetries, yielding execution times of 176.4 ms per view per OSC iteration for the largest dataset tested. It was also the fastest method overall for forward projection. The partial storage method was shown to be of low interest, yielding performance similar to the OTF approach, but adding the memory and geometry constraints of the fully stored method. In sum, the OTF ray-tracing seems to be an appropriate choice for general-purpose reconstruction libraries, while the fully pre-stored method may provide additional acceleration for specific CT systems with compatible scanner geometries and computing hardware. Further research on this topic is oriented towards hardware acceleration of multi-raytracing using half-precision (16-bit) floating-point arithmetic, which may allow for accurate simulation of a finite-size X-ray source and detector, to yield better image quality.

Acknowledgments

This work was supported in part by the Fonds de recherche du Québec—Nature et technologies (FRQ-NT). The authors acknowledge partial support by the CREATE Medical Physics Research Training Network grant of the Natural Sciences and Engineering Research Council of Canada (Grant No. 432290).

Disclosure of Conflicts of Interest

The authors have no relevant conflicts of interest to disclose.

Chapitre 5

Conclusion

La reconstruction itérative en tomodensitométrie est sans doute un domaine de recherche mature et ses retombées commencent à s'intégrer dans des produits commerciaux. Pour le bénéfice du patient, la RI permet une réduction appréciable de la dose de radiation ionisante grâce à la modélisation formelle de la statistique des photons et aux algorithmes de débruitage avancés. De plus, à dose égale, on observe une amélioration de la qualité d'image par rapport aux approches classiques via rétroprojection filtrée, assurant un diagnostic médical plus fiable. Les grands fabricants d'appareils de TDM pour la radiologie diagnostique ont récemment mis en oeuvre leurs algorithmes propriétaires, qui ont été évalués en termes de qualité d'image observée [8], mais dont les approches théoriques ne sont pas dans le domaine public. Par contre, les applications à plus petite échelle attendent encore des solutions taillées sur mesure. C'est le cas de divers appareils de TDM dédiés, comme les appareils de TDM à faisceau conique embarqués en radiothérapie, les appareils de TDM mobiles et les systèmes de tomographie optique. La présente thèse a contribué à l'amélioration des méthodes de reconstruction applicables à ces modalités d'imagerie. Les innovations proposées sont soutenues par une validation solide qui fait partie intégrale de ce travail de recherche.

5.1 Retour sur le travail accompli

Dans le cas de la TDM à rayons X à faisceau conique, l'algorithme de reconstruction OSC-TV a montré sa robustesse par rapport à une réduction de la dose de radiation ionisante, tout en conservant une qualité d'image suffisante pour des tâches comme la localisation de repères anatomiques. En tomographie optique, où la procédure d'imagerie n'est pas associée à des risques, OSC-TV a démontré une nette amélioration de la qualité d'image. Enfin, indépendamment de l'application, l'évaluation comparative des méthodes de calcul et de stockage de la matrice-système pour la géométrie à faisceau conique a permis de bien cerner les approches les plus rapides et les plus pratiques pour effectuer les opérations de projection et de rétroprojection en RI. Ce volet est essentiel pour l'élaboration de logiciels qui respectent les exigences

cliniques en termes de temps de reconstruction et exploitent de façon optimale le matériel informatique à la fine pointe de la technologie.

5.1.1 Retour sur l'algorithme de reconstruction itérative OSC-TV

L'algorithme de reconstruction itérative OSC-TV a été conçu pour l'imagerie basse-dose avant tout : la modélisation de l'atténuation des photons via la distribution de Poisson, ainsi que la régularisation ont démontré une bonne robustesse face au bruit associé à l'imagerie basse-dose. Les effets de la réduction de dose à 45% et à 25% de la dose standard ont été étudiés. Les propriétés de convergence de cet algorithme ont été explorées sur des données synthétiques et une analyse qualitative a été effectuée sur des données expérimentales. Plusieurs algorithmes semblables ont été proposés durant les dernières années ; toutefois, OSC-TV se distingue par un faible nombre de paramètres empiriques à ajuster pour différents protocoles d'acquisition, ce qui en fait un candidat compétitif pour le transfert de la technologie vers la clinique. Un autre aspect important est l'implantation de cet algorithme sur GPU. Les temps de reconstruction estimés sur une station de travail typique étant de plusieurs heures pour des images de taille clinique, OSC-TV a été d'emblée implanté sur GPU. Quelques stratégies d'accélération originales ont été élaborées dans le cadre de cette recherche, portant surtout sur l'utilisation judicieuse de la mémoire partagée et l'optimisation des accès à la mémoire globale du GPU, compte tenu que la reconstruction tomographique est avant tout limitée par la bande passante en raison de la complexité du modèle géométrique.

5.1.2 Retour sur l'évaluation de OSC-TV pour la tomographie optique

La tomographie optique est une modalité d'imagerie prometteuse utilisée en dosimétrie 3D pour la radiothérapie. Les propriétés des images reconstruites avec l'algorithme OSC-TV en tomographie optique à faisceau conique ont été évaluées dans le but de démontrer certains avantages de cette approche par rapport à la reconstruction classique. Comme l'imagerie optique est sans danger, la qualité du sinogramme est très élevée en comparaison avec le cas de la TDM à rayons X. Par ailleurs, très peu de publications ont été consacrées à la RI, et aucune ne présente de résultats obtenus à partir de données expérimentales. Ce volet de la recherche a donc contribué à combler cette lacune. Il a été montré que OSC-TV permet d'atteindre une résolution spatiale plus élevée et un bruit plus faible dans les images reconstruites. Ces métriques sont très importantes lors du contrôle de la qualité dosimétrique pour des faisceaux de radiothérapie conformationnelle, qui est caractérisée par des formes géométriques complexes et de forts gradients de dose. La réduction des artéfacts dus à des structures fortement atténuantes a aussi été observée. En plus de démontrer l'utilité de OSC-TV en tomographie optique, le travail de recherche a évalué le potentiel de la quantification absolue de la densité optique dans les images obtenues via tomographie optique à faisceau conique. Il a été démontré que les approches de reconstruction à modèle monochromatique sont limitées dans leur capacité à quantifier la densité optique, en raison d'un spectre étalé de la source lumineuse. Ainsi, un

algorithme qui tient compte des propriétés spectrales de la source, du milieu atténuant et du détecteur serait souhaitable pour ce type d'appareil.

5.1.3 Retour sur les méthodes de calcul et stockage de la matrice système

La reconstruction itérative est caractérisée par une complexité de calcul élevée, et cette dernière dépend beaucoup de la sophistication du modèle géométrique de l'intersection entre la radiation et le sujet imagé. Afin d'assurer des temps de reconstruction raisonnables pour d'éventuelles applications cliniques, le dernier volet du projet s'est penché sur la vitesse des opérateurs de projection et rétroprojection, utiles pour différents algorithmes itératifs, ainsi qu'à l'itération OSC. Quelques études récentes ont démontré qu'il est plus avantageux de stocker la matrice-système plutôt que de la recalculer en TDM hélicoïdale [87, 180]. Ces résultats ont été pris comme hypothèse de départ pour le cas de la TDM à faisceau conique avec une implantation GPU des opérateurs. Quelques approches de stockage des coefficients de la matrice-système ont été comparées au calcul à la volée des ces coefficients. Le modèle d'intersection de rayons minces avec des voxels sur une grille cartésienne, qui est une bonne représentation de la géométrie réelle, a été retenu pour cette étude. Trois approches ont été comparées : compression et stockage complet de la matrice-système, stockage partiel avec recalcul à la volée le cas échéant, ainsi que le traçage à la volée pur, sans recours à la mémoire globale du GPU. Chacune de ces méthodes a bénéficié de stratégies d'optimisation sur mesure. Pour les méthodes avec stockage, la compression via les symétries géométriques a été utilisée. Pour la méthode à stockage partiel, deux types de *kernel* différents ont été comparés, soit un kernel avec un processus léger par voxel et un processus léger par rayon. Pour toutes les méthodes, l'optimisation de la bande passante a été mise de l'avant via le concept de coalescence sur GPU. La méthode de stockage complet a été la plus rapide pour l'itération OSC, avec une accélération de 1.52× par rapport à la méthode avec traçage à la volée. Toutefois, cette méthode a des exigences très strictes sur la géométrie d'acquisition qui doit respecter plusieurs types de symétries, et nécessite une capacité élevée de la mémoire globale du GPU. La méthode de calcul à la volée a été de loin la plus rapide pour l'opérateur de projection directe. Les temps de rétroprojection et OSC ont été satisfaisants, et la méthode peut se vanter aussi de la plus grande flexibilité, car elle n'impose pas de propriétés géométriques spéciales à l'acquisition. Les méthodes de stockage partiel se sont montrées peu utiles, avec un temps d'exécution aussi élevé qu'avec le traçage à la volée et des restrictions géométriques de la méthode avec stockage complet. Au sens large, cette étude montre que pour l'implantation d'une librairie de reconstruction générique, là où la géométrie est difficile à prévoir, l'approche de traçage à la volée est à privilégier. Pour une application axée sur la vitesse de calcul, il serait d'intérêt de contraindre la géométrie d'acquisition et de profiter de l'accélération fournie par la méthode de stockage complet de la matrice-système.

5.2 Perspective

La présente recherche a donné lieu a un nouvel algorithme de reconstruction itératif en TDM implanté sur GPU, dont les propriétés en termes de qualité d'image résultante et de performance numérique sont d'intérêt pour le transfert vers des systèmes d'imagerie à faisceau conique commerciaux. De plus, le code source résultant de ces travaux est déjà repris comme plate-forme de développement de méthodes de reconstruction TDM itératives avec modélisation physique avancée. En marge de cette recherche, les premiers pas pour la modélisation et la quantification précise de la radiation diffusée ont été faits, avec une étude des facteurs diffusé/primaire calculés sur GPU en adaptant la plate-forme de simulation Monte-Carlo GPUMCD au contexte de la radiographie à rayons X [182]. Au sein du Groupe de recherche en physique médicale, les travaux se poursuivent sur l'intégration de la correction de la radiation diffusée à la méthode OSC-TV. De plus, l'utilisation de OSC-TV en tomodensitométrie dynamique (4D), avec une composante temporelle, a été étudiée au sein de ce groupe de recherche [183]. Enfin, l'algorithme est en phase d'évaluation pour d'autres applications connexes, soit l'imagerie non-destructive de spécimens géologiques, ainsi que la reconstruction d'images à très basse dose pour la tomodensitométrie dento-maxillo-faciale.

Quelques avenues intéressantes sont à explorer pour l'amélioration de la qualité d'image en TDM et la performance numérique du code à partir de la plate-forme OSC-TV. Entre autres, l'implantation de la régularisation par sous-images (NLM) [131, 132] en remplacement de la minimisation de la variation totale serait d'intérêt pour donner une apparence plus naturelle aux images reconstruites. L'implantation du modèle géométrique aux rayons fins multiples ou aux empreintes séparables serait un atout important pour la reconstruction d'images à très haute résolution spatiale. Du côté numérique, la conversion de certaines structures de données vers la précision de 16 bits [102] permettrait d'accélérer la reconstruction, car la bande passante du matériel serait mieux utilisée. Dans une perspective plus large, il faut être conscient que plusieurs méthodes de reconstruction itérative offrent une excellente convergence théorique. En revanche, au terme des travaux de recherche décrits dans cette thèse, OSC-TV se présente comme une méthode bien documentée et mature pour le transfert de technologie, ainsi qu'une plate-forme de développement pour des modèles physiques avancés. L'intégration de ces derniers constitue une priorité pour assurer le succès de la méthode dans les applications cliniques.

Annexe A

Spectrum flattening

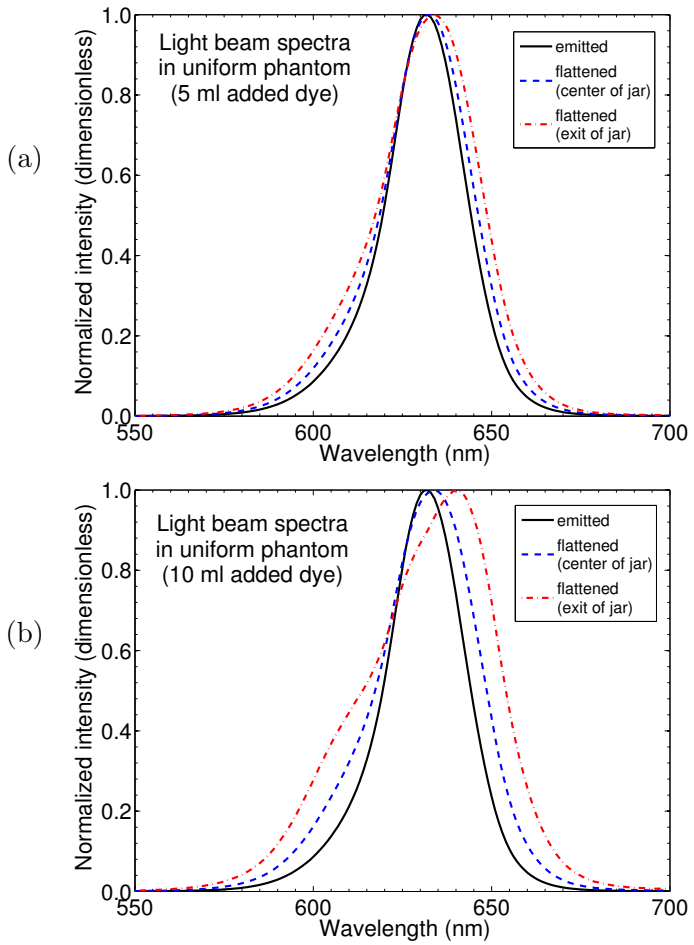


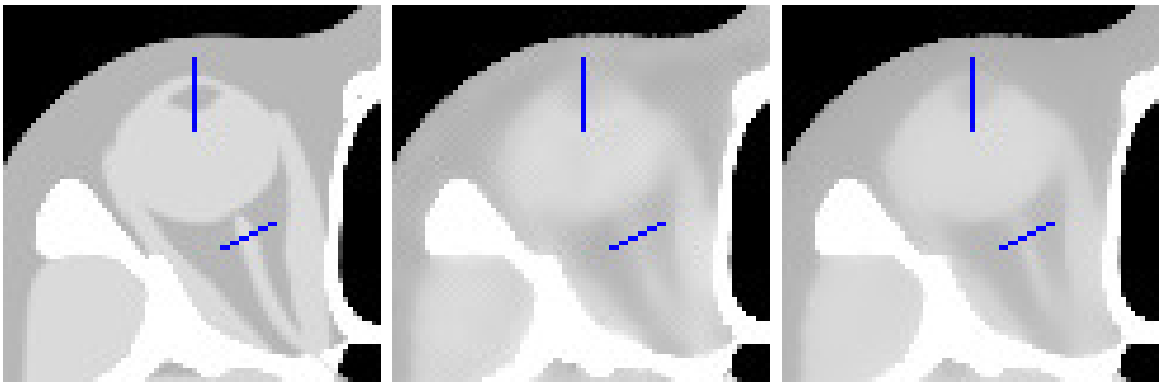
FIGURE A.1 – Simulated *spectrum flattening*, based on experimental DeskCAT red LEDs emission spectrum and dyed water absorption spectra for (a) 5.0 ml and (b) 10.0 ml added dye. Note the progressive expansion of the spectrum and its shift towards the red, more prominent for 10.0 ml added dye.

Annexe B

Reconstruction of low-contrast features by the OSC-TV algorithm

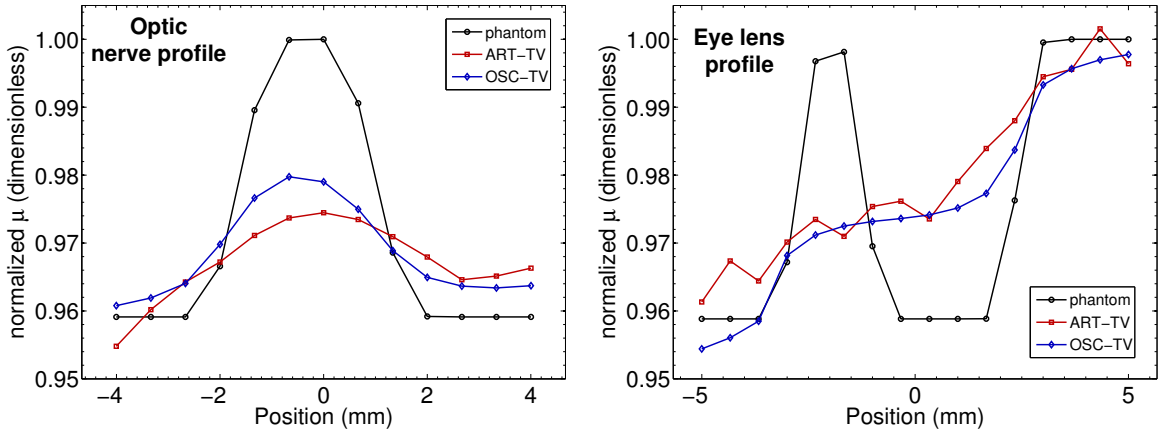
A numerical human head phantom was used to evaluate the over-regularization of small features of low relative contrast. The dataset was of a few-view type (200 projections) with Poisson noise to simulate a typical clinical cone-beam CT (CBCT) head scan, see Matenine *et al.* [168] for details. The OSC-TV algorithm was also compared to a well-known constrained projection onto convex sets algorithm described by Sidky *et al.* [56], denoted as ART-TV here. The optic nerve cross-section and the eye lens cross-section were analyzed, situated in a 256×256 mm^2 slice. The slices are presented in Fig. B.1 and the profiles, in Fig B.2. The FWHM's of the optic nerve profiles are the following : 2.6 mm, 3.8 mm and 2.9 mm respectively for the phantom, ART-TV and OSC-TV. Therefore, OSC-TV induces an 11% increase in the FWHM vs. 46% for ART-TV. The phantom relative contrast between the optic nerve and the surrounding tissues is of +4.3%, while ART-TV and OSC-TV reconstruct respectively +1.3% and +1.8% contrast, both demonstrating important bias. The main conclusion stemming from these results is that the OSC-TV method is spatially edge-preserving, but small structures suffer from contrast loss.

For the eye lens, the simplified phantom profile (Fig. B.2 (b)) has multiple structures, from left to right : the cornea (approximated as muscle), lens (approximated as fat) and vitreous humor (as muscle). The FWHM of the cornea is of 1.5 mm ; it was not properly reconstructed, and seems to have biased the neighbouring lens μ value. In this case, the cornea and lens may be interpreted as a line pair and are averaged as one structure by both algorithms, apparently reaching the spatial resolution limit.



(a) Phantom (b) ART-TV (c) OSC-TV

FIGURE B.1 – Synthetic head phantom slice, right eye region shown (ROI : 80×80 voxels or 53.3×53.3 mm^2). The displayed relative μ interval is of $[0.51, 0.72]$ with 1.0 being the phantom bone density. (a) Phantom (b) ART-TV reconstruction (c) OSC-TV reconstruction. The blue lines represent acquired μ profiles across the optic nerve and the eye lens. Note excessive blurring of edges for both algorithms, with ART-TV also suffering from residual noise.



(a) (b)

FIGURE B.2 – Small features' profiles : (a) across the optic nerve and (b) across the eye lens. For the optic nerve, the reconstructed profiles are broader and of lower contrast. For the eye lens, the small features are lost. The μ values are normalized to the maximum of the phantom profile.

Annexe C

Handling Geometrical Symmetries

Geometrical symmetries may be present in the projection dataset and the formalism presented here uses a decision logic which allows for mixed symmetry levels to be exploited. A few important definitions are the following : consider the source-detector axis be a particular line segment that originates at the source focal point and joins the detector surface at a normal incidence. Let this point on the detector be denoted *source axis contact point*. The possible detector displacements or *shifts* are defined with regard to this point.

In this work, the requirements for using symmetries were tested on floating-point numeric values of the parameters. Therefore, user-supplied tolerances were used to determine whether the conditions are met. Values expected to be nonzero were tested to be equal up to a relative error of 0.1%. Values expected to be nil were tested based on the same error, relative to a small reference non-zero measurement, such as a reasonable angle increment between projections or a voxel side length.

The most likely symmetry, and also the easiest to apply, is the z-axis mirror symmetry. The source axis contact point should fall in the middle of a detector row or on the boundary between two detector rows. The current code simply requires that the detector shift in the z-direction be nil, to avoid complex partial and conditional reuse of system matrix data. Also, the source-detector axis should cross the boundary between two voxels or cross the middle of a voxel (in the z-direction) when entering the reconstructed volume.

Another likely symmetry is the repetition of ray patterns in the different quadrants of the gantry rotation plane. The first set of conditions to use this type of symmetry concerns the reconstructed volume. The voxels should be square in the gantry rotation plane and the reconstructed volume should be square and centered in the gantry rotation plane. The starting angle of the acquisition is unimportant in this case, but the angular increment $\Delta\phi$ should be such that an integer number of *angular-reusable projection angles* n_a covers an integer number

N_q of 90° (quadrant) intervals :

$$n_a \approx \frac{N_q \times 90^\circ}{\Delta\phi}, \{N_q, n_a\} \subset \mathbb{N} \tag{C.1}$$

The most interesting case is when $N_q = 1$, and the partial system matrix may be reused at every 90° interval, as shown in Fig. C.1 (a). For multi-revolution scans, even $N_q = 4$ and beyond may be used for system matrix data reuse. Using this formalism, incomplete (sub-revolution) scans may also benefit from data reuse when $\Delta\phi$ is compliant.

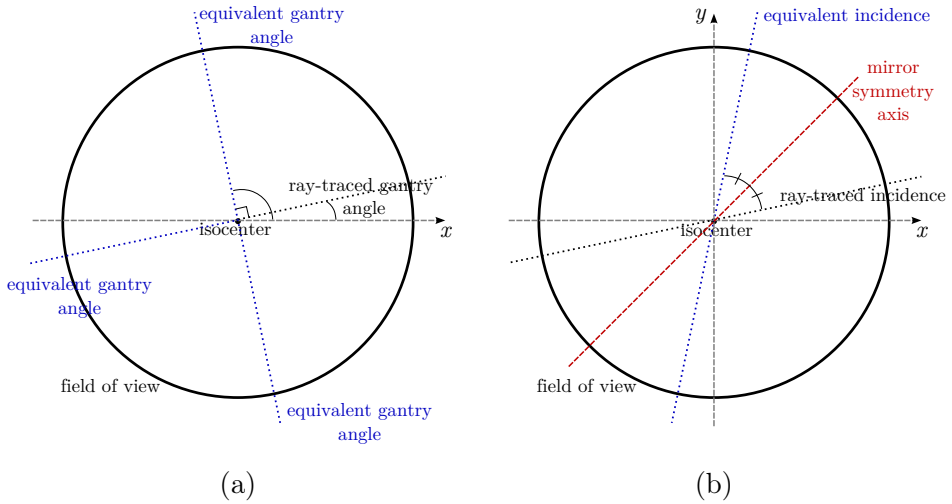


FIGURE C.1 – Geometrical symmetries. (a) With a compatible angular increment, up to 4 equivalent incidence angles arise in the acquisition ; voxel references may be recomputed via rotation. (b) With a compatible angular increment and starting angle, equivalent incidences arise in the]45°,90°] interval ; voxel references may be recomputed via mirror symmetry along the $\phi = 45^\circ$ axis.

The final symmetry is the mirror symmetry along the $\phi = 45^\circ$ radial plane, in the 1st quadrant. It allows for reusing the data from the $[0, 45]^\circ$ range in the $]45, 90]^\circ$ range, see Fig. C.1 (b). It imposes several conditions, so it is much less likely to be applicable. The conditions required for the rotation symmetry must be met, and supplementary conditions appear. The initial acquisition angle relative to the volume coordinate system must be a multiple of 90°. The previously mentioned N_q must be 1, and the detector shift in the direction tangent to the gantry rotation trajectory must be a multiple of the half-width of a detector bin. The current code simply requires that the detector shift in the tangent direction be 0, to avoid complex partial and conditional reuse of system matrix data.

Bibliographie

[1] J.M. Boone, W.R. Hendee, M.F. McNitt-Gray, and S.E. Seltzer. Radiation exposure from CT scans : how to close our knowledge gaps, monitor and safeguard exposure—proceedings and recommendations of the Radiation Dose Summit, sponsored by NIBIB, February 24–25, 2011. *Radiology*, 265(2) :544–554, 2012.

[2] M.D. Cohen. ALARA, Image Gently and CT-induced cancer. *Pediatric Radiology*, 45(4) :465–470, 2015.

[3] ICRP. Managing patient dose in computed tomography. *Annals of the ICRP*, 30(4) :7, 2000.

[4] L. Feldkamp, L. Davis, and J.W. Kress. Practical cone-beam algorithm. *J. Opt. Soc. Amer. A*, 1 :612–619, June 1984.

[5] J. Bian, J.H. Siewerdsen, X. Han, E.Y. Sidky, J.L. Prince, C.A. Pelizzari, and X. Pan. Evaluation of sparse-view reconstruction from flat-panel-detector cone-beam CT. *Physics in Medicine and Biology*, 55(22) :6575, 2010.

[6] M. Beister, D. Kolditz, and W.A. Kalender. Iterative reconstruction methods in X-ray CT. *Physica Medica*, 28(2) :94 – 108, 2012.

[7] J. Nuyts, B.D. Man, J.A. Fessler, W. Zbijewski, and F.J. Beekman. Modelling the physics in the iterative reconstruction for transmission computed tomography. *Physics in Medicine and Biology*, 58(12) :R63, 2013.

[8] A. Omotayo and I. Elbakri. Objective performance assessment of five computed tomography iterative reconstruction algorithms. *Journal of X-Ray Science and Technology*, 24(6) :913–930, 2016.

[9] J.T. Bushberg, J. Bert, E.M. Leidholdt, Jr., and J.M. Boone. *The essential physics of medical imaging, 2nd edition.* Lippincott Williams And Wilkins, 2002.

[10] R. Ledley, J. Wilson, T. Golab, and L. Rotolo. The ACTA-Scanner : The whole body computerized transaxial tomograph. *Computers in Biology and Medicine*, 4(2) :145–155, 1974.

[11] H.E. Johns and J.R. Cunningham. *The Physics of Radiology, fourth edition.* Charles C Thomas, Springfield IL, 1983.

[12] J. Zhang, V. Weir, L. Fajardo, J. Lin, H. Hsiung, and E.R. Ritenour. Dosimetric characterization of a cone-beam O-arm™ imaging system. *Journal of X-ray Science and Technology*, 17(4) :305–317, 2009.

[13] K. Abramovitch and D.D. Rice. Basic Principles of Cone Beam Computed Tomography. *Dental Clinics of North America*, 58(3) :463 – 484, 2014.

[14] E.L. Ritman. Current status of developments and applications of micro-CT. *Annual review of biomedical engineering*, 13 :531–552, 2011.

[15] E.B. Podgorsak, editor. *Radiation Oncology Physics : A Handbook for Teachers And Students.* International Atomic Energy Agency, 2003.

[16] D. Jaffray, J.H. Siewerdsen, J.W. Wong, and A.A. Martinez. Flat-panel cone-beam computed tomography for image-guided radiation therapy. *International Journal of Radiation Oncology*Biology*Physics*, 53(5) :1337–1349, 2002.

[17] J.P. Bissonnette, P.A. Balter, L. Dong, K.M. Langen, D.M. Lovelock, M. Miften, D.J. Moseley, J. Pouliot, J.J. Sonke, and S. Yoo. Quality assurance for image-guided radiation therapy utilizing CT-based technologies : A report of the AAPM TG-179. *Medical Physics*, 39(4) :1946–1963, 2012.

[18] X. Pan, E.Y. Sidky, and M. Vannier. Why do commercial CT scanners still employ traditional, filtered back-projection for image reconstruction ? *Inverse Problems*, 25(12) :123009, 2009.

[19] E.Y. Sidky, Y. Duchin, X. Pan, and C. Ullberg. A constrained, total-variation minimization algorithm for low-intensity x-ray CT. *Medical Physics*, 38(S1) :S117–S125, 2011.

[20] X. Jia, Y. Lou, J. Lewis, R. Li, X. Gu, C. Men, W.Y. Song, and S.B. Jiang. GPU-based fast low-dose cone beam CT reconstruction via total variation. *Journal of X-ray science and technology*, 19(2) :139–154, 2011.

[21] X. Jia, B. Dong, Y. Lou, and S.B. Jiang. GPU-based iterative cone-beam CT reconstruction using tight frame regularization. *Physics in Medicine and Biology*, 56(13) :3787, 2011.

[22] D. Stsepankou, A. Arns, S.K. Ng, P. Zygmanski, and J. Hesser. Evaluation of robustness of maximum likelihood cone-beam CT reconstruction with total variation regularization. *Physics in Medicine and Biology*, 57(19) :5955, 2012.

[23] H. Yan, X. Wang, F. Shi, T. Bai, M. Folkerts, L. Cervino, S.B. Jiang, and X. Jia. Towards the clinical implementation of iterative low-dose cone-beam CT reconstruction

in image-guided radiation therapy : Cone/ring artifact correction and multiple GPU implementation. *Medical Physics*, 41(11), 2014.

[24] X. Jia and S. Jiang, editors. *Graphics Processing Unit-Based High Performance Computing in Radiation Therapy*. CRC Press, 2015.

[25] K. Jordan. Advances in optical CT scanning for gel dosimetry. *Journal of Physics : Conference Series*, 3(1) :115, 2004.

[26] S. Doran and N. Krstajic. The history and principles of optical computed tomography for scanning 3-D radiation dosimeters. *Journal of Physics : Conference Series*, 56(1) :45, 2006.

[27] S. Doran. The history and principles of optical computed tomography for scanning 3-D radiation dosimeters : 2008 update. *Journal of Physics : Conference Series*, 164(1) :012020, 2009.

[28] L.J. Schreiner. Where does gel dosimetry fit in the clinic ? *Journal of Physics : Conference Series*, 164(1) :012001, 2009.

[29] M. Oldham, J.H. Siewerdsen, S. Kumar, J. Wong, and D. Jaffray. Optical-CT gel-dosimetry I : Basic investigations. *Medical Physics*, 30(4) :623–634, 2003.

[30] P. Shrimpton, B.F. Wall, and E.S. Fisher. The tissue-equivalence of the Alderson Rando anthropomorphic phantom for X-rays of diagnostic qualities. *Physics in Medicine and Biology*, 26(1) :133, 1981.

[31] K. Jessen, P. Shrimpton, J. Geleijns, W. Panzer, and G. Tosi. Dosimetry for optimisation of patient protection in computed tomography. *Applied Radiation and Isotopes*, 50(1) :165–172, 1999.

[32] C. Constantinou, F.H. Attix, and B.R. Paliwal. A solid water phantom material for radiotherapy x-ray and gamma-ray beam calibrations. *Medical Physics*, 9(3) :436–441, 1982.

[33] L. Beaulieu, M. Goulet, L. Archambault, and S. Beddar. Current status of scintillation dosimetry for megavoltage beams. In *Journal of Physics : Conference Series*, volume 444, page 012013, June 2013.

[34] L.J. Schreiner. Review of Fricke gel dosimeters. *Journal of Physics : Conference Series*, 3(1) :9, 2004.

[35] C. Baldock, Y.D. Deene, S. Doran, G. Ibbott, A. Jirasek, M. Lepage, K.B. McAuley, M. Oldham, and L.J. Schreiner. Polymer gel dosimetry. *Physics in Medicine and Biology*, 55(5) :R1, 2010.

[36] A. Appleby, A. Leghrouz, and E. Christman. Radiation chemical and magnetic resonance studies of aqueous agarose gels containing ferrous ions. *International Journal of*

Radiation Applications and Instrumentation. Part C. Radiation Physics and Chemistry, 32(2) :241 – 244, 1988.

[37] S. Doran and D.N.B. Yatigammana. Eliminating the need for refractive index matching in optical CT scanners for radiotherapy dosimetry : I. Concept and simulations. *Physics in Medicine and Biology*, 57(3) :665, 2012.

[38] K. Chisholm, D. Miles, L. Rankine, and M. Oldham. Investigations into the feasibility of optical-CT 3D dosimetry with minimal use of refractively matched fluids. *Medical Physics*, 42(5) :2607–2614, 2015.

[39] S. Bosi, P. Naseri, A. Puran, J. Davies, and C. Baldock. Initial investigation of a novel light-scattering gel phantom for evaluation of optical CT scanners for radiotherapy gel dosimetry. *Physics in Medicine and Biology*, 52(10) :2893, 2007.

[40] T. Olding, O. Holmes, and L.J. Schreiner. Cone beam optical computed tomography for gel dosimetry I : scanner characterization. *Physics in Medicine and Biology*, 55(10) :2819, 2010.

[41] T. Olding and L.J. Schreiner. Cone-beam optical computed tomography for gel dosimetry II : imaging protocols. *Physics in Medicine and Biology*, 56(5) :1259–1279, 2011.

[42] J.L. Prince and J. Links. *Medical Imaging Signals and Systems*. Pearson Prentice Hall, 2006.

[43] Y. Kyriakou, M. Meyer, and W.A. Kalender. Technical note : Comparing coherent and incoherent scatter effects for cone-beam CT. *Physics in Medicine and Biology*, 53(10) :N175–N185, 2008.

[44] X. Jia, H. Yan, L. Cervi no, M. Folkerts, and S.B. Jiang. A GPU tool for efficient, accurate, and realistic simulation of cone beam CT projections. *Medical physics*, 39(12) :7368–7378, 2012.

[45] M. Slaney and A. Kak. *Principles of computerized tomographic imaging*. IEEE Press, 1988.

[46] A. Macovski. *Medical Imaging Systems*. Prentice-Hall, 1983.

[47] G. Pratx and L. Xing. GPU computing in medical physics : A review. *Medical Physics*, 38(5) :2685–2697, 2011.

[48] R. Gordon, R. Bender, and G.T. Herman. Algebraic Reconstruction Techniques (ART) for three-dimensional electron microscopy and X-ray photography. *Journal of Theoretical Biology*, 29(3) :471 – 481, 1970.

[49] G.N. Hounsfield. Computerized transverse axial scanning (tomography). 1. Description of system. *Br J Radiol*, 46(552) :1016–1022, December 1973.

[50] A. Andersen and A. Kak. Simultaneous Algebraic Reconstruction Technique (SART) : A superior implementation of the ART algorithm. *Ultrasonic Imaging*, 6(1) :81 – 94, 1984.

[51] K. Lange and R. Carson. EM reconstruction algorithms for emission and transmission tomography. *Journal of Computer Assisted Tomography*, 8(2) :306, 1984.

[52] K. Lange. Overview of Bayesian methods in image reconstruction. In A.F. Gmitro, P.S. Idell, and I.J. LaHaie, editors, *Digital Image Synthesis and Inverse Optics*, volume 1351, pages 270–287. SPIE, July 1990.

[53] K. Lange and J.A. Fessler. Globally convergent algorithms for maximum a posteriori transmission tomography. *Image Processing, IEEE Transactions on*, 4(10) :1430–1438, October 1995.

[54] C. Kamphuis and F.J. Beekman. Accelerated iterative transmission CT reconstruction using an ordered subsets convex algorithm. *Medical Imaging, IEEE Transactions on*, 17(6) :1101–1105, December 1998.

[55] S. Ahn, J.A. Fessler, D. Blatt, and A. Hero. Convergent incremental optimization transfer algorithms : application to tomography. *Medical Imaging, IEEE Transactions on*, 25(3) :283–296, 2006.

[56] E.Y. Sidky, C.M. Kao, and X. Pan. Accurate image reconstruction from few-views and limited-angle data in divergent-beam CT. *X-ray Sci. Tech.*, 14(2) :119–139, 2006.

[57] E.Y. Sidky and X. Pan. Image reconstruction in circular cone-beam computed tomography by constrained, total-variation minimization. *Physics in Medicine and Biology*, 53(17) :4777, 2008.

[58] J. Wang, T. Li, and L. Xing. Iterative image reconstruction for CBCT using edge-preserving prior. *Medical Physics*, 36(1) :252–260, 2009.

[59] Z. Yu, J.B. Thibault, C.A. Bouman, K.D. Sauer, and J. Hsieh. Fast Model-Based X-Ray CT Reconstruction Using Spatially Nonhomogeneous ICD Optimization. *IEEE Transactions on Image Processing*, 20(1) :161–175, January 2011.

[60] M.G. McGaffin and J.A. Fessler. Alternating Dual Updates Algorithm for X-ray CT Reconstruction on the GPU. *IEEE Transactions on Computational Imaging*, 1(3) :186–199, 2015.

[61] J.F. Bonnans, J.C. Gilbert, C. Lemaréchal, and C.A. Sagastizábal. *Numerical optimization : theoretical and practical aspects*. Springer Science And Business Media, 2006.

[62] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1) :1–38, 1977.

[63] K. Sauer and C. Bouman. A local update strategy for iterative reconstruction from projections. *Signal Processing, IEEE Transactions on*, 41(2) :534 –548, February 1993.

[64] R.H. Byrd, P. Lu, J. Nocedal, and C. Zhu. A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM Journal on Scientific Computing*, 16(5) :1190–1208, 1995.

[65] C. Zhu, R.H. Byrd, P. Lu, and J. Nocedal. Algorithm 778 : L-BFGS-B : Fortran Subroutines for Large-scale Bound-constrained Optimization. *ACM Trans. Math. Softw.*, 23(4) :550–560, December 1997.

[66] B. Hamelin, Y. Goussard, and J.P. Dussault. Penalized-likelihood region-of-interest CT reconstruction by local object supersampling. In *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 739–742, August 2007.

[67] Y. Goussard, M. Golkar, A. Wagner, and M. Voorons. Cylindrical coordinate representation for statistical 3D CT reconstruction. In *Proceedings of the 13th International Meeting on Fully Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine*, pages 138–141, 16-21 June 2013.

[68] C. Thibaudeau, J.D. Leroux, R. Fontaine, and R. Lecomte. Fully 3D iterative CT reconstruction using polar coordinates. *Medical Physics*, 40(11) :111904, 2013.

[69] A. Soriano, M.J. Rodríguez-Alvarez, A. Iborra, F. Sanchez, M. Carles, P. Conde, A.J. González, L. Hernández, L. Moliner, A. Orero, L.F. Vidal, and J.M. Benlloch. EM tomographic image reconstruction using polar voxels. *Journal of Instrumentation*, 8(01) :C01004, 2013.

[70] R.M. Lewitt. Alternatives to voxels for image representation in iterative reconstruction algorithms. *Physics in Medicine and Biology*, 37(3) :705, 1992.

[71] S. Matej and R.M. Lewitt. Practical considerations for 3-D image reconstruction using spherically symmetric volume elements. *IEEE Transactions on Medical Imaging*, 15(1) :68–78, February 1996.

[72] A. Ziegler, T. Köhler, T. Nielsen, and R. Proksa. Efficient projection and backprojection scheme for spherically symmetric basis functions in divergent beam geometry. *Medical Physics*, 33(12) :4653–4663, 2006.

[73] F. Xu and K. Mueller. A comparative study of popular interpolation and integration methods for use in computed tomography. In *Biomedical Imaging : Nano to Macro, 2006. 3rd IEEE International Symposium on*, pages 1252–1255, 2006.

[74] J. Hsieh. *Computed Tomography, Second Edition : Principles, Design, Artifacts, and Recent Advances*. SPIE Press, 2009. DOI : 10.1117/3.817303 eISBN : 9780819480422.

[75] J.G. Donaire and I. García. A comparison of several interpolation methods in 3D X-ray cone beam reconstruction. In *9th European Signal Processing Conference (EUSIPCO 1998)*, pages 1–4, 1998.

[76] P.M. Joseph. An Improved Algorithm for Reprojecting Rays through Pixel Images. *IEEE Transactions on Medical Imaging*, 1(3) :192–196, November 1982.

[77] B. De Man and S. Basu. Distance-driven projection and backprojection in three dimensions. *Physics in Medicine and Biology*, 49(11) :2463, 2004.

[78] Y. Long, J.A. Fessler, and J. Balter. 3D Forward and Back-Projection for X-Ray CT Using Separable Footprints. *IEEE Transactions on Medical Imaging*, 29(11) :1839–1850, November 2010.

[79] R.L. Siddon. Fast calculation of the exact radiological path for a three-dimensional CT array. *Medical Physics*, 12(2) :252–255, 1985.

[80] F. Jacobs, E. Sundermann, B.D. Sutter, M. Christiaens, and I. Lemahieu. A Fast Algorithm to Calculate the Exact Radiological Path Through a Pixel Or Voxel Space. *Journal of Computing and Information Technology*, 6 :89–94, 1998.

[81] M. Christiaens, B.D. Sutter, K.D. Bosschere, J.V. Campenhout, and I. Lemahieu. A fast, cache-aware algorithm for the calculation of radiological paths exploiting subword parallelism. *J. Syst. Archit.*, 45(10) :781–790, 1999.

[82] Z. Chen, R. Ning, and D.L. Conover. Accurate perspective projection calculation using a pixel-pyramid model for iterative cone beam reconstruction. In *Proc. SPIE*, volume 5030, pages 728–739, 2003.

[83] Z. Chen and R. Ning. Pixel-pyramid model for divergent projection geometry. *Optical Engineering*, 44(2) :027002–027002–10, 2005.

[84] J.A. Browne, J.M. Boone, and T.J. Holmes. Maximum-likelihood x-ray computed-tomography finite-beamwidth considerations. *Appl. Opt.*, 34(23) :5199–5209, August 1995.

[85] H. Zhao and A.J. Reader. Fast ray-tracing technique to calculate line integral paths in voxel arrays. In *Nuclear Science Symposium Conference Record, 2003 IEEE*, volume 4, pages 2808–2812, October 2003.

[86] F. Xu. Fast Implementation of Iterative Reconstruction with Exact Ray-Driven Projector on GPUs. *Tsinghua Science And Technology*, 15(1) :30 – 35, 2010.

[87] J. Xu and B.M.W. Tsui. Iterative image reconstruction in helical cone-beam x-ray CT using a stored system matrix approach. *Physics in Medicine and Biology*, 57(11) :3477, 2012.

[88] C. Jian-lin, Z. Han-ming, Y. Bin, L. Lei, G. Ming, and W. Lin-yuan. Matrix approach for processing of iterative reconstruction on cone beam CT. In *2013 IEEE International Conference on Medical Imaging Physics and Engineering*, pages 72–77, October 2013.

[89] J. Dongarra. Trends in high performance computing : a historical overview and examination of future developments. *IEEE Circuits and Devices Magazine*, 22(1) :22–27, January 2006.

[90] J.C. Cuevas and E. Scheer. *Molecular electronics : an introduction to theory and experiment*. World Scientific, 2010.

[91] K. Asanovic, R. Bodik, J. Demmel, T. Keaveny, K. Keutzer, J. Kubiatowicz, N. Morgan, D. Patterson, K. Sen, J. Wawrzynek, D. Wessel, and K. Yelick. A View of the Parallel Computing Landscape. *Commun. ACM*, 52(10) :56–67, October 2009.

[92] NVIDIA Corporation Inc., Santa Clara CA. *NVIDIA CUDA Reference Manual version 3.1*, 2010.

[93] J.E. Stone, D. Gohara, and G. Shi. OpenCL : A parallel programming standard for heterogeneous computing systems. *Computing in science and engineering*, 12(1-3) :66–73, 2010.

[94] R. Couturier. *Designing Scientific Applications on GPUs*. Chapman And Hall/CRC, 2013. ISBN 1466571624, 9781466571624.

[95] NVIDIA Corporation Inc., Santa Clara CA. *CUDA C PROGRAMMING GUIDE version 7.5*, September 2015.

[96] B. Barney. Introduction to Parallel Computing. Lawrence Livermore National Laboratory, 2012.

[97] G.M. Amdahl. Validity of the Single Processor Approach to Achieving Large Scale Computing Capabilities. In *Proceedings of the April 18-20, 1967, Spring Joint Computer Conference*, AFIPS '67 (Spring), pages 483–485, New York, NY, USA, 1967. ACM.

[98] MATLAB. *version 9.1 (R2016b)*. The MathWorks Inc., Natick, Massachusetts, 2016.

[99] K. Karimi, N.G. Dickson, and F. Hamze. A performance comparison of CUDA and OpenCL. *arXiv preprint arXiv :1005.2581*, 2010.

[100] J. Hoberock and N. Bell. Thrust : A Parallel Template Library, 2010. Version 1.3.0.

[101] M.A. Nassiri, S. Hissoiny, J.F. Carrier, and P. Després. Fast GPU-Based Computation of the Sensitivity Matrix for a PET List-Mode OSEM Algorithm. *Physics in Medicine and Biology*, 57(19) :6279, 2012.

[102] C. Maaß, M. Baer, and M. Kachelrieß. CT image reconstruction with half precision floating-point values. *Medical Physics*, 38(S1) :S95–S105, 2011.

[103] J.S. Kole and F.J. Beekman. Evaluation of the ordered subset convex algorithm for cone-beam CT. *Phys. Med. Biol.*, 50(4) :613–623, 2005.

[104] J. Xu and B.M.W. Tsui. A Compound Poisson Maximum-Likelihood Iterative Reconstruction Algorithm for X-Ray CT. In *9th International Meeting on Fully Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine*, pages 108–111, 2007.

[105] M. Yan, J. Chen, L. Vese, J. Villasenor, A. Bui, and J. Cong. EM+TV Based Reconstruction for Cone-Beam CT with Reduced Radiation. In G. Bebis, R. Boyle, B. Parvin, D. Koracin, S. Wang, K. Kyungnam, B. Benes, K. Moreland, C. Borst, S. DiVerdi, C. Yi-Jen, and J. Ming, editors, *Advances in Visual Computing*, volume 6938 of *Lecture Notes in Computer Science*, pages 1–10. Springer Berlin / Heidelberg, 2011.

[106] R.E. Alvarez and A. Macovski. Energy-selective reconstructions in X-ray computerized tomography. *Physics in Medicine and Biology*, 21 :733–744, 1976.

[107] B. De Man, J. Nuyts, P. Dupont, G. Marchal, and P. Suetens. An iterative maximum-likelihood polychromatic algorithm for CT. *Medical Imaging, IEEE Transactions on*, 20(10) :999 –1008, October 2001.

[108] M. Kachelrieß and W.A. Kalender. Improving PET/CT attenuation correction with iterative CT beam hardening correction. In *Nuclear Science Symposium Conference Record, 2005 IEEE*, volume 4, pages 1905–1909, October 2005.

[109] Y. Kyriakou, E. Meyer, D. Prell, and M. Kachelriess. Empirical beam hardening correction (EBHC) for CT. *Medical Physics*, 37(10) :5179–5187, 2010.

[110] C. Lemmens, D. Faul, and J. Nuyts. Suppression of Metal Artifacts in CT Using a Reconstruction Procedure That Combines MAP and Projection Completion. *IEEE Transactions on Medical Imaging*, 28(2) :250–260, February 2009.

[111] E. Meyer, R. Raupach, M. Lell, B. Schmidt, and M. Kachelrieß. Normalized metal artifact reduction (NMAR) in computed tomography. *Medical Physics*, 37(10) :5482–5493, 2010.

[112] A. Souza. Fast-forward projection approach for 3D iterative metal artifact suppression. In *Proc. SPIE*, volume 8313, pages 83133P–83133P–6, 2012.

[113] M. Stille, M. Kleine, J. Hägele, J. Barkhausen, and T.M. Buzug. Augmented Likelihood Image Reconstruction. *IEEE Transactions on Medical Imaging*, 35(1) :158–173, January 2016.

[114] E.P. Rührnschopf and K. Klingenbeck. A general framework and review of scatter correction methods in x-ray cone-beam computerized tomography. Part 1 : Scatter compensation approaches. *Medical Physics*, 38(7) :4296–4311, 2011.

[115] E.P. Rührnschopf and and K. Klingenbeck. A general framework and review of scatter correction methods in cone beam CT. Part 2 : Scatter estimation approaches. *Medical Physics*, 38(9) :5186–5199, 2011.

[116] D. Donoho. Compressed Sensing. *Information Theory, IEEE Transactions on,*, 52(4) :1289 –1306, April 2006.

[117] J. Dong-Jiang, H. Zhang, and Z. Xiao-Bing. TV OS-SART with Fractional Order Integral Filtering. In *2012 Eighth International Conference on Computational Intelligence and Security*, pages 132–135, November 2012.

[118] Y. Du, X. Wang, X. Xiang, and Z. Wei. Evaluation of hybrid SART + OS + TV iterative reconstruction algorithm for optical-CT gel dosimeter imaging. *Physics in Medicine and Biology*, 61(24) :8425, 2016.

[119] W. Li-yan and W. Zhi-hui. Fast gradient-based algorithm for total variation regularized tomography reconstruction. In *2011 4th International Congress on Image and Signal Processing*, volume 3, pages 1572–1576, October 2011.

[120] B. Song, J.C. Park, and W.Y. Song. A low-complexity 2-point step size gradient projection method with selective function evaluations for smoothed total variation based CBCT reconstructions. *Physics in Medicine and Biology*, 59(21) :6565, 2014.

[121] Z. Tian, X. Jia, K. Yuan, T. Pan, and S.B. Jiang. Low-dose CT reconstruction via edge-preserving total variation regularization. *Physics in Medicine and Biology*, 56(18) :5949, 2011.

[122] Y. Liu, J. Ma, Y. Fan, and Z. Liang. Adaptive-weighted total variation minimization for sparse data toward low-dose x-ray computed tomography image reconstruction. *Physics in Medicine and Biology*, 57(23) :7923, 2012.

[123] M. Debatin, P. Zygmanski, D. Stsepankou, and J. Hesser. CT reconstruction from few-views by Anisotropic Total Variation minimization. In *2012 IEEE Nuclear Science Symposium and Medical Imaging Conference Record (NSS/MIC)*, pages 2295–2296, October 2012.

[124] M. Lee, Y. Han, J.P. Ward, M. Unser, and J.C. Ye. Interior Tomography Using 1D Generalized Total Variation. Part II : Multiscale Implementation. *SIAM Journal on Imaging Sciences*, 8(4) :2452–2486, 2015.

[125] J. Chen, L. Wang, B. Yan, H. Zhang, and G. Cheng. Efficient and robust 3D CT image reconstruction based on total generalized variation regularization using the alternating direction method. *Journal of X-ray science and technology*, 23(6) :683–699, 2015.

[126] N. Sun, T. Sun, J. Wang, and S. Tan. CBCT reconstruction via a penalty combining total variation and its higher-degree term. In *Proc. SPIE*, volume 9412, pages 94123T–94123T–9, 2015.

[127] C.W. Seo, B.K. Cha, S. Jeon, Y. Huh, J.C. Park, B. Lee, J. Baek, and E. Kim. Compressed sensing with gradient total variation for low-dose CBCT reconstruction. *Nuclear Instruments and Methods in Physics Research Section A : Accelerators, Spectrometers, Detectors and Associated Equipment*, 784 :570 – 573, 2015.

[128] E.Y. Sidky, R. Chartrand, J.M. Boone, and X. Pan. Constrained TpV Minimization for Enhanced Exploitation of Gradient Sparsity : Application to CT Image Reconstruction. *IEEE Journal of Translational Engineering in Health and Medicine*, 2 :1–18, 2014.

[129] A. Cai, L. Wang, B. Yan, L. Li, H. Zhang, and G. Hu. Efficient TpV minimization for circular, cone-beam computed tomography reconstruction via non-convex optimization. *Computerized Medical Imaging and Graphics*, 45 :1 – 10, 2015.

[130] H. Kim, J. Chen, A. Wang, C. Chuang, M. Held, and J. Pouliot. Non-local total-variation (NLTV) minimization combined with reweighted L1-norm for compressed sensing CT reconstruction. *Physics in Medicine and Biology*, 61(18) :6878, 2016.

[131] A. Buades, B. Coll, and J.M. Morel. A review of image denoising algorithms, with a new one. *Multiscale Modeling and Simulation*, 4(2) :490–530, 2005.

[132] A. Buades, B. Coll, and J.M. Morel. Non-Local Means Denoising. *Image Processing On Line*, 1 :208–212, 2011.

[133] Z. Zheng, E. Papenhausen, and K. Mueller. DQS advisor : a visual interface and knowledge-based system to balance dose, quality, and reconstruction speed in iterative CT reconstruction with application to NLM-regularization. *Physics in Medicine and Biology*, 58(21) :7857, 2013.

[134] X. Jia, Z. Tian, Y. Lou, J.J. Sonke, and B. Jiang. Four-dimensional cone beam CT reconstruction and enhancement using a temporal nonlocal means method. *Medical Physics*, 39(9) :5592–5602, September 2012.

[135] D. Kim, S. Ramani, and J.A. Fessler. Combining Ordered Subsets and Momentum for Accelerated X-Ray CT Image Reconstruction. *IEEE Transactions on Medical Imaging*, 34(1) :167–178, January 2015.

[136] W. Yao and K. Leszczynski. Analytically derived weighting factors for transmission tomography cone beam projections. *Physics in Medicine and Biology*, 54(3) :513, 2009.

[137] J. Brokish, D.B. Keesing, and Y. Bresler. Iterative circular conebeam CT reconstruction using fast hierarchical backprojection/reprojection operators. In *Proc. SPIE*, volume 7622, pages 76221R–76221R–9, 2010.

[138] J. Brokish, P. Sack, and Y. Bresler. Combined algorithmic and GPU acceleration for ultra-fast circular conebeam backprojection. In *Proc. SPIE*, volume 7622, pages 762256–762256–9, 2010.

[139] C. Miao, B. Liu, Q. Xu, and H. Yu. An improved distance-driven method for projection and backprojection. *Journal of X-ray science and technology*, 22(1) :1–18, 2014.

[140] A. Mitra, D.G. Politte, B.R. Whiting, J.F. Williamson, and J.A. O'Sullivan. Multi-GPU Acceleration of Branchless Distance Driven Projection and Backprojection for Clinical Helical CT. *Journal of Imaging Science and Technology*, 2016.

[141] D. Schlifske and H. Medeiros. A fast GPU-based approach to branchless distance-driven projection and back-projection in cone beam CT. In *Proc. SPIE*, volume 9783, pages 97832W–97832W–8, 2016.

[142] J. Dittmann. Efficient ray tracing on 3D regular grids for fast generation of digitally reconstructed radiographs in iterative tomographic reconstruction techniques. *arXiv preprint arXiv :1609.00958*, 2016.

[143] D. Karimi and R. Ward. On the computational implementation of forward and back-projection operations for cone-beam computed tomography. *Medical And Biological Engineering And Computing*, 54(8) :1193–1204, 2016.

[144] K. Hahn, H. Schöndube, K. Stierstorfer, J. Hornegger, and F. Noo. A comparison of linear interpolation models for iterative CT reconstruction. *Medical Physics*, 43(12) :6455–6473, 2016.

[145] C.Y. Chou, Y.Y. Chuo, Y. Hung, and W. Wang. A fast forward projection using multithreads for multirays on GPUs in medical image reconstruction. *Medical Physics*, 38(7) :4052–4065, 2011.

[146] H. Gao. Fast parallel algorithms for the x-ray transform and its adjoint. *Medical physics*, 39(11) :7110–7120, 2012.

[147] V.G. Nguyen and S.J. Lee. Parallelizing a Matched Pair of Ray-Tracing Projector and Backprojector for Iterative Cone-Beam CT Reconstruction. *IEEE Transactions on Nuclear Science*, 62(1) :171–181, February 2015.

[148] L. Rankine and M. Oldham. On the feasibility of optical-CT imaging in media of different refractive index. *Medical Physics*, 40(5), 2013.

[149] P. Gilbert. Iterative methods for the three-dimensional reconstruction of an object from projections. *Journal of Theoretical Biology*, 36(1) :105 – 117, 1972.

[150] Y. Du, X. Wang, and X. Xiang. Artifacts suppression in optical CT for gel dosimeters by iterative reconstruction. In *Journal of Physics : Conference Series*, volume 573, page 012063. IOP Publishing, 2015.

[151] M.W. Kan, L.H. Leung, W. Wong, and N. Lam. Radiation Dose From Cone Beam Computed Tomography for Image-Guided Radiation Therapy. *International Journal of Radiation Oncology*Biology*Physics*, 70(1) :272 – 279, 2008.

[152] K. Lange, M. Bahn, and R. Little. A Theoretical Study of Some Maximum Likelihood Algorithms for Emission and Transmission Tomography. *Medical Imaging, IEEE Transactions on*, 6(2) :106–114, June 1987.

[153] F.J. Beekman and C. Kamphuis. Ordered subset reconstruction for x-ray CT. *Physics in medicine and biology*, 46(7) :1835–1844, 2001.

[154] T. Zinsser and B. Keck. Systematic Performance Optimization of Cone-Beam Back-Projection on the Kepler Architecture. In *Proceedings of the 13th International Meeting on Fully Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine*, 16-21 June 2013.

[155] J.S. Kole and F.J. Beekman. Parallel statistical image reconstruction for cone-beam x-ray CT on a shared memory computation platform. *Phys. Med. Biol.*, 50(6) :1265–1272, 2005.

[156] M. Debatin, D. Stsepankou, and J. Hesser. CT Reconstruction from Few-Views by Higher Order Adaptive Weighted Total Variation Minimization. In *Proceedings of the 13th International Meeting on Fully Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine*, 16-21 June 2013.

[157] D. Bertsekas. A New Class of Incremental Gradient Methods for Least Squares Problems. *SIAM Journal on Optimization*, 7(4) :913–926, 1997.

[158] W.P. Segars, M. Mahesh, T.J. Beck, E.C. Frey, and B.M.W. Tsui. Realistic CT simulation using the 4D XCAT phantom. *Medical Physics*, 35(8) :3800–3808, 2008.

[159] J. Kaipio and E. Somersalo. Statistical inverse problems : Discretization, model reduction and inverse crimes. *Journal of Computational and Applied Mathematics*, 198(2) :493 – 504, 2007.

[160] Å. Palm, E. Nilsson, and L. Herrnsdorf. Absorbed dose and dose rate using the Varian OBI 1.3 and 1.4 CBCT system. *Journal of Applied Clinical Medical Physics*, 11(1) :229–240, 2010.

[161] J. Ma, Z. Liang, Y. Fan, Y. Liu, J. Huang, W. Chen, and H. Lu. Variance analysis of x-ray CT sinograms in the presence of electronic noise background. *Medical Physics*, 39(7) :4051–4065, 2012.

[162] D. Matenine. Reconstruction itérative sur matériel graphique en tomodensitométrie. Master's thesis, Université de Montréal, 2011.

[163] J. Šolc and V. Sochor. Feasibility of radiochromic gels for 3D dosimetry of brachytherapy sources. *Metrologia*, 49(5) :S231–S236, 2012.

[164] S. Babic, J. Battista, and K. Jordan. Three-Dimensional Dose Verification for Intensity-Modulated Radiation Therapy in the Radiological Physics Centre Head-and-Neck Phan-

tom Using Optical Computed Tomography Scans of Ferrous Xylenol-Orange Gel Dosimeters. *International Journal of Radiation Oncology*Biology*Physics*, 70(4) :1281 – 1291, 2008.

[165] H.S. Sakhalkar and M. Oldham. Fast, high-resolution 3D dosimetry utilizing a novel optical-CT scanner incorporating tertiary telecentric collimation. *Med. Phys.*, 35(1) :101 – 111, 2008.

[166] N. Krstajic and S. Doran. Characterization of a parallel-beam CCD optical-CT apparatus for 3D radiation dosimetry. *Physics in Medicine and Biology*, 52(13) :3693, 2007.

[167] W.G. Campbell, D.A. Rudko, N.A. Braam, D.M. Wells, and A. Jirasek. A prototype fan-beam optical CT scanner for 3D dosimetry. *Medical Physics*, 40(6) :061712, 2013.

[168] D. Matenine, J. Mascolo-Fortin, Y. Goussard, and P. Després. Evaluation of the OSC-TV Iterative Reconstruction Algorithm for Cone-Beam Optical CT. *Medical Physics*, 42(11) :6376–6386, 13 October 2015.

[169] M. Oldham, J.H. Siewerdsen, A. Shetty, and D. Jaffray. High resolution gel-dosimetry by optical-CT and MR scanning. *Medical Physics*, 28(7) :1436–1445, 2001.

[170] S.M. Bentzen. Evaluation of the spatial resolution of a CT scanner by direct analysis of the edge response function. *Medical Physics*, 10(5) :579–581, 1983.

[171] D. Ebbing and S.D. Gammon. *General chemistry.* Cengage Learning, 9th edition, 2010.

[172] Y.D. Deene. Computational simulations of the influence of noise in optical CT reconstruction. *Journal of Physics : Conference Series*, 573(1) :12076–12079, 2015.

[173] D. Jaffray and J.H. Siewerdsen. Cone-beam computed tomography with a flat-panel imager : Initial performance characterization. *Medical physics*, 27(6) :1311–1323, 2000.

[174] W. Zhuang, S.S. Gopal, and T.J. Hebert. Numerical evaluation of methods for computing tomographic projections. *IEEE Transactions on Nuclear Science*, 41(4) :1660–1665, August 1994.

[175] F. Xu, W. Xu, M. Jones, B. Keszthelyi, J. Sedat, D. Agard, and K. Mueller. On the efficiency of iterative ordered subset reconstruction algorithms for acceleration on GPUs. *Computer Methods and Programs in Biomedicine*, 98(3) :261 – 270, 2010.

[176] S. Ha, A. Kumar, and K. Mueller. A Study of Volume Integration Models for Iterative Cone-Beam Computed Tomography. In *Proceedings of the 13th Meeting on Fully 3D Image Reconstruction*, pages 464–467, 2015.

[177] S. Ha, H. Li, and K. Mueller. Efficient area-based ray integration using summed area tables and regression models. In *Proceedings of the 4th International Meeting on image formation in X-ray CT*, pages 507–510, 2016.

[178] G.L. Zeng and G.T. Gullberg. Unmatched projector/backprojector pairs in an iterative reconstruction algorithm. *IEEE Transactions on Medical Imaging*, 19(5) :548–555, May 2000.

[179] P. Després, J. Rinkel, B. Hasegawa, and S. Prevrhal. Stream processors : a new platform for Monte Carlo calculations. In F. Verhaegen, editor, *Journal of Physics : Conference Series*, volume 102, page 012007, 2008.

[180] M. Guo and H. Gao. Memory-Efficient Algorithm for Stored Projection and Backprojection Matrix in Helical CT. *Medical Physics*, 2017.

[181] D. Matenine, Y. Goussard, and P. Després. GPU-accelerated regularized iterative reconstruction for few-view cone beam CT. *Medical Physics*, 42(4) :1505–1517, 13 March 2015.

[182] D. Matenine, Y. Goussard, and P. Després. Fast Monte-Carlo simulation of Cone-Beam X-ray image formation using GPUMCD. International Workshop on Monte Carlo Techniques in Medical Physics, 17-20 June 2014. Quebec City, QC.

[183] J. Mascolo-Fortin, D. Matenine, and P. Després. Adaptation of the OSC-TV Reconstruction Algorithm for 4D Cone Beam Computed Tomography. In M. Kachelrieß, editor, *Proceedings of the 4th International Conference on Image Formation in X-Ray Computed Tomography*, 18-22 July 2016.