UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Metal Artifact Reduction in Computed Tomography

Permalink

https://escholarship.org/uc/item/9w55m91h

Author

Karimi, Seemeen

Publication Date

2014

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Metal Artifact Reduction in Computed Tomography

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Electrical Engineering (Signal and Image Processing)

by

Seemeen Karimi

Committee in charge:

Pamela Cosman, Chair
William Hodgkiss
Xiaoqian Jiang
Harry Martz
Truong Nguyen
Nuno Vasconcelos

2014

Copyright
Seemeen Karimi, 2014
All rights reserved.

The dissertation of Seemeen Karimi is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

_____
Chair

University of California, San Diego

2014

DEDICATION

For Fareeda and Sophie

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I am most grateful to my family from the start. My parents made very many sacrifices for my education. My husband, Aziz, provided mentorship, positivity and top-notch IT support. He is the first and last checkpoint of everything I do, and I cannot adequately express here how much he has done for me. I am thankful to Sophie, my little daughter, for her acceptance of my long hours, and for her excitement and pride when I would have a paper accepted or do well in a course. She is the most understanding person I have seen. And to Pierre and Gaston for keeping me company while I worked over the years. Pierre and Gaston are cats.

I am very grateful to my academic advisor, Prof. Pamela Cosman, for giving me the opportunity to be her student. She advised me with kindness, fairness, wisdom and patience. Prof. Cosman is enormously committed to the success and well-being of her students. She always made time for me despite a senior professor's schedule. She shared her insights, made me question my assumptions, taught me how to write better, and encouraged me.

Thanks to Dr. Harry Martz, my co-advisor at Lawrence Livermore National Laboratories, for sponsoring my research. Dr. Martz took a chance on me and supported my research even when it diverged from his initial vision. I am grateful for his ideas, his humor, his confidence in me, and for painstakingly reviewing drafts and catching embarrassing mistakes.

Thanks to my committee for their participation and recommendations on my research. In particular I am deeply grateful to Dr. Xiaoqian Jiang for hours of discussions in which he helped me generate and refine ideas, and for the use of his computers. His wholehearted support of this work was indispensable. I thank Prof. Vasconcelos for his patience with me in teaching me what research is about and what things are worth thinking about. Thanks also to Profs. Hodgkiss and Nguyen, for gently encouraging me to think, which made the research stronger. Thanks also to Dr. Stanley Chan for his optimization methods and generously sharing his time and his ideas, and to Jihoon Kim for help in statistical tests.

I thank Dr. Chris Wald for sharing medical data, which required him to spend quite a lot of his time. Without data, I would have been quite stuck. I

thank Dr. Carl Crawford for encouraging me to enter the PhD program although I was afraid to do it, and for training at a previous job. I am deeply grateful to my friend Karan Sikka for the occasional chats that always yielded something useful. I am also grateful to Pengchong Jin for the use of forward projection software and to Dr. Jeff Kallman for helpful discussions. This research was funded by Lawrence Livermore National Laboratories, through the Awareness and Localization of Explosives program of the Department of Homeland Security, Science and Technology Directorate.

Chapter 2 contains material from the paper "Segmentation of Artifacts and Anatomy in CT Metal Artifact Reduction," *Medical Physics*, Vol. 39, Issue 10, 2012. This paper was co-authored by Pamela Cosman, Harry Martz and Christoph Wald. Chapter 3 contains material from "Metal Artifact Reduction for CT based Luggage Screening", which was co-authored by Pamela Cosman and Harry Martz and was presented at the *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014*. Chapter 5 contains material from "Flexible Methods for Segmentation Evaluation: Results from CT Luggage Screening," *Journal of X-Ray Science and Technology*, Vol. 22, Issue 2, 2014. This was co-authored by Xiaoqian Jiang, Pamela Cosman and Harry Martz.

VITA

1995	B. E. in Biomedical Engineering, Mumbai University
1997	M. S. in Biomedical Engineering, University of North Carolina, Chapel Hill
1997-2004	Analogic Corporation, Boston
2004-2013	NeuroLogica Corporation, Boston
2010-2014	Graduate Research Assistant, University of California, San Diego
2014	Ph. D. in Electrical Engineering (Signal and Image Processing), University of California, San Diego

PUBLICATIONS

S. Karimi, P.C. Cosman, H. Martz, and C. Wald, "Using Segmentation in CT Metal Artifact Reduction", *IEEE Southwest Symposium on Image Analysis and Interpretation, SSIAI*, 2012

S. Karimi, X. Jiang, P.C. Cosman, and H. Martz, "Evaluation of Segmentation Algorithms in CT Scanning", *2nd IEEE Conference on Healthcare Informatics, Imaging, and Systems Biology*, 2012. Best Poster Award.

S. Karimi, H. Martz, and P.C. Cosman, "Metal Artifact Reduction for CT-Based Luggage Screening", *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2014

S. Karimi, P. Cosman, C. Wald, and H. Martz, "Segmentation of Artifacts and Anatomy in CT Metal Artifact Reduction," *Medical Physics*, Vol. 39, Issue 10, 5857-5868, 2012

S. Karimi, X. Jiang, P. Cosman, and H. Martz, "Flexible Methods for Segmentation Evaluation: Results from CT Luggage Screening," *Journal of X-Ray Science and Technology*, Vol. 22, Issue 2, 175-195, 2014

ABSTRACT OF THE DISSERTATION

Metal Artifact Reduction in Computed Tomography

by

Seemeen Karimi

Doctor of Philosophy in Electrical Engineering (Signal and Image Processing)

University of California, San Diego, 2014

Pamela Cosman, Chair

In computed tomography (CT) imaging, if metal is present in the scan, it gives rise to streaks and shadows called metal artifacts. We consider two applications of CT, radiology and luggage screening for aviation security. In radiology, metal artifacts make it difficult to evaluate anatomical structures. In luggage screening, computations on metal-artifact degraded images give rise to false alarms. Therefore metal artifact reduction (MAR) is an active area of research.

For medical imaging, we improve upon a class of MAR algorithms that are often called sinogram completion methods. The sinogram (Radon transform) contains the log-attenuation measured by the scanner. In sinogram completion methods, portions of the sinogram contaminated by metal are replaced with estimates of the underlying data. Our improvement comes from segmenting artifacts

from anatomy, based on their spatial and intensity distributions. Segmentation yields an intermediate image which when forward-projected, guides the sinogram completion. The corrected sinogram is reconstructed into the final image. We applied our algorithm to CT scans of the head and found that our results improved upon the state-of-the-art.

In luggage screening, the variety of scanned articles is larger and the amount of metal is greater, therefore assumptions cannot be made on spatial and intensity distributions. Our strategy here is a hybrid one, combining numerical optimization with sinogram completion. The numerical optimization de-emphasizes metal-contaminated projections. We compared our method to previously published MAR algorithms qualitatively and quantitatively. Our method reduces metal artifacts and preserves more image details than the compared methods.

We also developed methods to evaluate the accuracy of segmentation algorithms in CT. The first method is based on mutual information of machine segments (MS) against ground truth (GT) segments. Mutual information is computed from a confusion matrix that contains the quantity of a feature common to MS and GT labels. The second method is based on feature recovery. We compute optimal one-to-one correspondence between GT and MS labels, and extract total and systematic errors. The errors give us insights that can be used for improving the algorithms. The evaluation of these methods themselves was based on synthetic problems and human observer evaluation.

Chapter 1

Introduction

Computed Tomography (CT) is an imaging modality used in radiology, aviation security and other applications. A CT scanner measures the attenuation of x-rays by materials, and CT images represent linear attenuation coefficients at each point. In medical imaging, different biological tissues have different attenuation of x-rays, which appear as different levels of brightness in a CT image. The CT images are reviewed by a radiologist for structural or physiological findings. In aviation security, passenger luggage is scanned in explosives detection systems (EDS), in which a CT scanner generates images. The images are analyzed by automatic target recognition (ATR) algorithms.

ATR includes segmentation of potential threats. Segmentation is a challenging task for intrinsic and extrinsic reasons. Chemically different materials have overlapping CT density ranges, and there is a nearly unlimited variety of articles in luggage. This task is intrinsically difficult even if CT images were perfect. In addition, the ATR algorithms must contend with image degradation from metal objects. When metal is present in CT scans, it generates streaks and shadows in images, called metal artifacts, that obscure the surrounding data. For both reasons, in luggage screening, ordinary non-threat objects may produce false alarms. Resolving false alarms involves high labor cost because alarm bags must be unpacked or sent for secondary screening [1]. This makes lowering the false alarm rate an important goal in security scanning.

This thesis is focused on two research problems. The first problem is metal

artifact reduction (MAR). The problem of metal artifacts exists in medical imaging, which is the most common use of CT. The artifacts obscure information about anatomical structures, making it difficult for radiologists to correctly interpret the images or for computer programs to analyze them. Most of the MAR research is in medical imaging. We started our investigation there. Highly effective MAR can be achieved by making assumptions on the contents of the scan. Here the greater emphasis is on image quality, and less on how general it is. However, in luggage, no assumptions can be made about the contents of the scan. Here the MAR methods must be robust. The different needs in these applications encourage different approaches to MAR.

The second problem we focus on is the evaluation of segmentation algorithms. As noted above, the segmentation of luggage images is a difficult task. The accuracy of segmentation algorithms must be meaningfully quantified in order to evaluate them. Insights into the behavior of the algorithms is desirable so that they can be tuned or improved.

In this chapter, we present some of the fundamental concepts in CT scanning, which are necessary for this research. Then we introduce the problem of metal artifacts in detail, including causes, simulations, and a literature review. Finally, we explain the challenges of segmentation evaluation.

1.1 CT data and image reconstruction

In this section, we give a brief explanation of how cross-sectional images are reconstructed from CT scanner data, and the underlying assumptions of the reconstruction methods that relate to our research. Several texts have an excellent detailed treatment of CT image reconstruction [2–4].

At its heart, a CT scanner consists of one or more x-ray sources that generate a collimated polyenergetic x-ray beam, as depicted in Fig. 1.1. The figure depicts a third generation CT scanner, which is the most commonly used. The x-rays are attenuated by the scanned objects in their path, and then detected by an array of x-ray detectors. The detectors are scintillator crystal or ceramic elements

of a few millimeters in area. The information about the attenuation paths lies in the discrepancy between the emitted and detected x-rays. The x-ray source and detector assemblies rotate around an axis, to capture data from various angular positions. The axis is perpendicular to the plane of rotation, and the intersection of this axis and the plane is called the isocenter, marked with a "+" in the picture. The patient bed (or bag conveyor) translates along the same axis, allowing multiple 2D cross-sections to be imaged in a step-and-shoot mode or in a volume scanning mode. The set of data collected in this manner is called a sinogram or projection data. The projection data are reconstructed into a CT image.

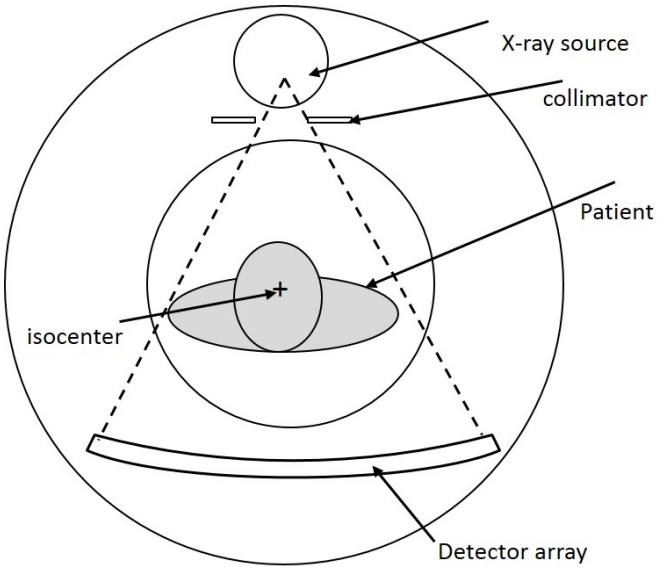


Figure 1.1: A simplified diagram of a CT scanner.

1.1.1 Analytical reconstruction

Consider a function $f(M)$ where M is a vector representing a point in some coordinate system. Consider lines or planes $\mathcal{L}$ with normal vector $\overrightarrow{n}$ and distance ρ from the origin. The Radon transform is an integral transform [3].

$$Rf(\rho, \overrightarrow{n}) = \int_{M \in \mathcal{L}(\rho, \overrightarrow{n})} f(M) dM \tag{1.1}$$

Therefore, it is a linear transform. In two dimensions, we define

$$p(\theta, t) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\delta(x\cos\theta + y\sin\theta - t)dxdy \tag{1.2}$$

CT scanner projection data are usually acquired in fan beams. The rays in fan beam data converge at the source as shown in Fig. 1.1. The fan rays are usually sorted into a set of parallel rays in a procedure called re-binning [2]. These parallel projections are the 2D Radon transform, also known as a sinogram. The parallel projections are reconstructed into a two-dmensional image. A CT image along with its sinogram is shown in Fig. 1.2.

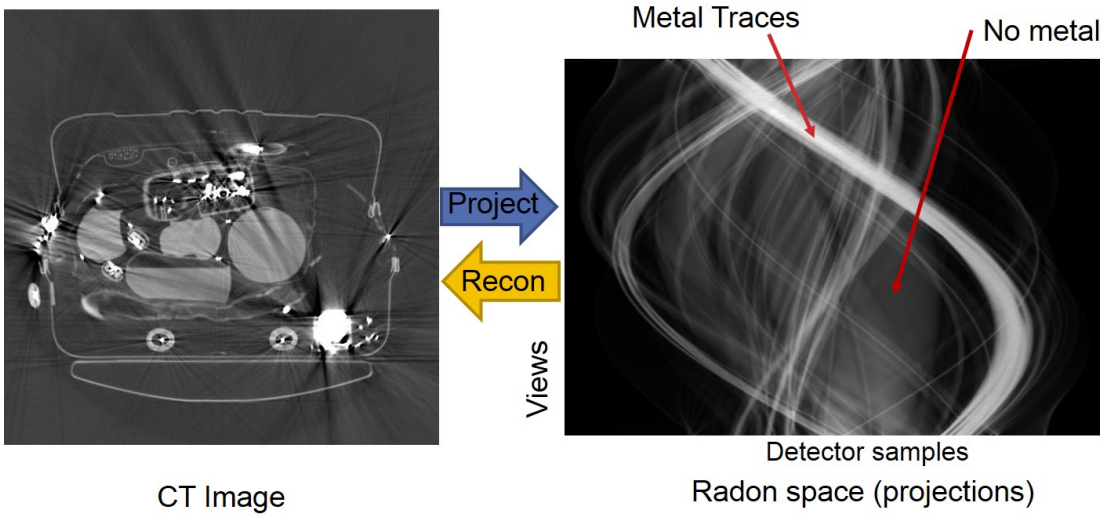


Figure 1.2: A CT image and projection data (Sinogram). In the sinogram, the view direction is vertical and the detector samples in a view are along the x-axis.

Taking the Fourier transform (FT) of the parallel projections, Eq. 1.2 gives

$$\int_{-\infty}^{\infty} p(\theta, t)\exp^{-2\pi j\omega t}\, dt = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\delta(x\cos\theta + y\sin\theta - t)e^{-2\pi j\omega t}dxdydt \tag{1.3}$$

$$P_\theta(\omega) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)e^{-2\pi j\omega(x\cos\theta + y\sin\theta)}dxdy \tag{1.4}$$

$$\begin{aligned} P_\theta(\omega) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)e^{-2\pi j\omega(\omega\cos\theta x + \omega\cos\theta y)}dxdy \\ &= F(\omega\cos\theta, \omega\sin\theta) \end{aligned} \tag{1.5}$$

This equation is known as the Fourier Slice Theorem, and is the basis for analytical tomographic reconstruction. It says that the 2D FT along a radial line is equal to the 1D FT of the projection data at the view angle equal to that of the radial line. Therefore it provides the required relationship between projections and the image through the FT.

Now considering

$$f(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F(u,v)e^{2\pi j(ux+vy)}dudv \tag{1.6}$$

We change to polar co-ordinates

$$f(x,y) = \int_{-\pi}^{\pi}\int_{0}^{\infty} \tilde{F}(\omega,\theta)e^{2\pi j(\omega\cos\theta x+\omega\sin\theta y)}\omega d\omega d\theta \tag{1.7}$$

Using the relationship $\tilde{F}(\theta,\omega) = \tilde{F}(\theta+\pi,-\omega)$, we get

$$\begin{aligned} f(x,y) &= \int_{0}^{\pi}\int_{-\infty}^{\infty} |\omega|\tilde{F}(\omega,\theta)e^{2\pi j\omega(\cos\theta x+\sin\theta y)}d\omega d\theta \\ &= \int_{0}^{\pi}\int_{-\infty}^{\infty} |\omega|P_\theta(\omega)e^{2\pi j\omega(x\cos\theta+y\sin\theta)}d\omega d\theta \end{aligned} \tag{1.8}$$

This equation, known as filtered backprojection (FBP), is the most commonly used form of Radon transform inversion in CT. The inner integral consists of filtering with the "Ram-Lak" kernel, and the outer integral is backprojection. Since this is a transform inversion, it assumes ideal data. Since the data from a CT scanner are not ideal, they are preprocessed in various ways to approximate ideal (and consistent) data. The preprocessing includes corrections for offset (dark current), gain, faulty-detector elements, cross-talk, after-glow, beam hardening, scatter and detector element-specific nonlinearities. Details of the complete practical system are outside the scope of this thesis.

1.1.2 Statistical reconstruction

Statistical reconstruction algorithms model the non-ideal nature of data acquisition. These algorithms can model geometric imperfections or irregular acquisition geometry and, most importantly, include noise models. A large body of literature has been developed. Well known reconstruction algorithms include

maximum likelihood expectation maximization (MLEM) [5], maximum a posteriori (MAP) reconstruction [6], and Algebraic Reconstruction Technique (ART) [7]. Since these algorithms include a noise model, they can obtain lower-noise reconstructions than analytical reconstruction, which makes them useful for low-dose scanning. These algorithms are slow, and not yet in commercial use.

An inverse problem can be expressed by the model $Ax = b$. In statistical reconstruction, the system matrix A is the forward projector, b is the observed data (projection data) and x are the model parameters we must estimate. MLEM gives the maximum likelihood estimates for the parameters x. MAP reconstruction allows the use of a prior, which provides regularization. The prior is usually a Gaussian or generalized Gaussian pdf. ART is a projection-onto-convex-sets approach that is fast, but does not model noise or natively use regularization.

1.1.3 Image units and viewing

Reconstructed images are represented in Hounsfield units (HU). In the HU scale, water is represented by 0 HU, and air by -1000 HU. Since material attenuation depends on x-ray energy, this scale is an artificial one. CT scanners must be calibrated so that air and water take on those values. Sometimes, a modified Hounsfield unit is used (MHU). The MHU scale is offset from the HU scale so that air is zero and water is 1000 MHU. In industrial CT, units of inverse centimeters are sometimes used.

The display tools for digital images usually show about 256 gray-scale levels. A viewing window width (WW) maps a range of CT values onto the display range. The center of the window, called the window level (WL), is at the center of the display range.

1.2 Metal artifacts: causes and previous research

1.2.1 Sources of metal artifacts

Metal artifacts are caused by beam hardening (the preferential attenuation of low energy photons in a polyenergetic x-ray beam), photon scatter, partial volume effects, photon starvation, and data sampling errors [2, 8]. Data sampling errors can be caused by inexact detector or view positions, cone beam effects [8] or patient motion. Streaks and shadows are created in the images, degrading their quality for viewing or processing [2]. Beam hardening and scatter may cause large, deep artifacts, that cannot be removed by filtering or destreaking methods. An image with metal artifacts is shown in Fig. 1.3.

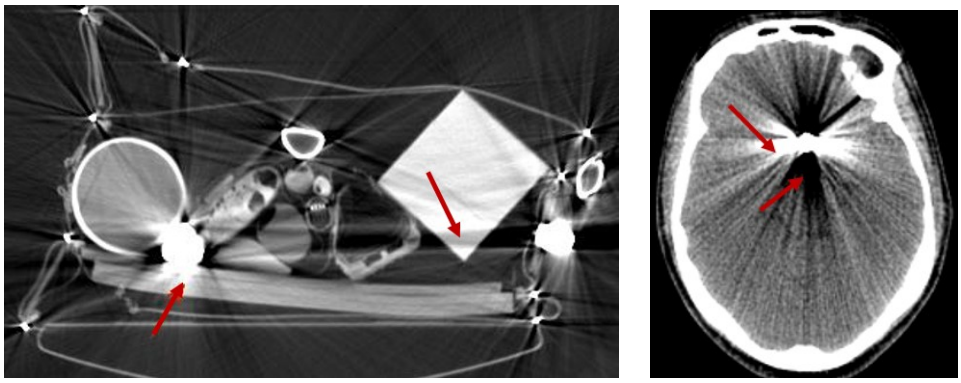


Figure 1.3: A CT image of a suitcase (left) and CT image of a head (right) have metal inserts. The artifacts from the metal inserts are indicated by the arrows.

In a linear reconstruction method, the underlying assumption is that the data are ideal and x-rays are monochromatic. Suppose I_0 represents the number of photons emitted from the x-ray tube per unit time, along a particular direction, and I represents the number of photons at the detector. Then monoenergetic attenuation is described by the Beer-Lambert Law [4]:

$$I(s,\theta) = I_0 e^{-\int_{s+l\theta \in L} \mu(s+l\theta, E_0) dl} \tag{1.9}$$

The integration here is over the scanned space L between source and detector, s is a vector representing the source position, and θ is a unit vector in the direction from the source position to a detector element. The Radon space data, i.e., projections,

are recovered by

$$p(s,\theta) = \int_{s+l\theta \in L} \mu(s+l\theta, E_0)dl = \log\left(\frac{I_0}{I(s,\theta)}\right) \tag{1.10}$$

In CT scanners, the x-ray beam is polyenergetic. Fig. 1.4 shows an example of an x-ray spectrum.

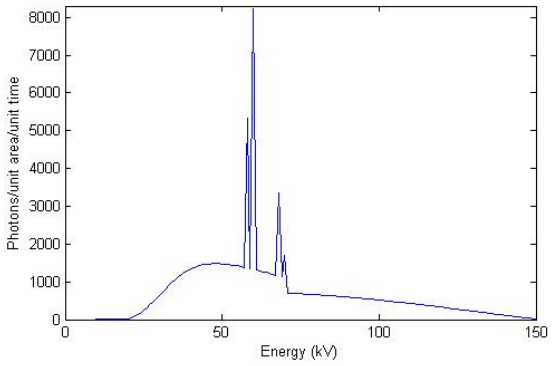


Figure 1.4: A spectrum from an x-ray tube.

If the spectrum is denoted by $S(E)$, then polyenergetic attenuation is described by the following equation [4]:

$$I(s,\theta) = \int S(E)e^{-\int_{s+l\theta \in L} \mu(s+l\theta,E)dl} dE \tag{1.11}$$

This equation means that the attenuation measured with Eq. (1.10) always underestimates the true attenuation, as shown below.

CT scanners are calibrated to scan tissue-like substances so that images can be reconstructed with linear algorithms. This procedure is called "flattening". The calibration consists of scanning a water phantom and measuring the apparent attenuation for different lengths of the phantom. A polynomial is fit to the measured attenuation and a compensating polynomial is calculated. This polynomial correction is applied to all the measured data. Then, the corrected projections can be reconstructed by FBP or other linear algorithms. Tissue like substances are similar to water, and the polynomial correction gives an acceptable approximation to the data.

When metals are present in the scans, the polynomial approximation does not suffice to correct the data. The measured projections become inconsistent, and

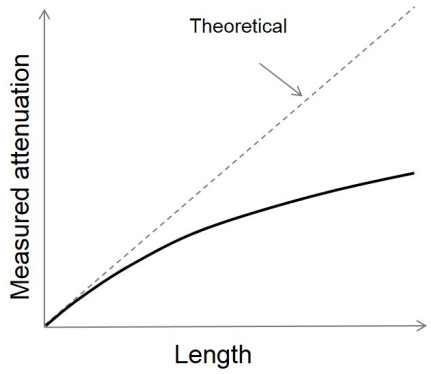


Figure 1.5: Theoretical and measured attenuation against object length. The measured attenuation is always lower than the theoretical attenuation, due to beam hardening.

the assumption of linearity in reconstruction breaks down, giving rise to artifacts. As we will see through the use of simulations in the next section, simple calibrations cannot be accurately performed for metals, as are done with water.

1.2.2 Simulation study

In order to understand how beam hardening from metals appears as artifacts, we simulated axial projections of ellipsoids in air of varying eccentricity, using the spectrum of Fig. 1.4. The spectrum was obtained using the program XSPECW2 [9]. The tube voltage is 150 kVp, the beam filtration is 4 mm of Aluminum, and there is a 10 degree Tungsten anode target. The simulated material is iron, whose energy-dependent attenuation cross-sections we obtain with XCOM [10]. There are 1400 projections, the detector spacing is 0.5 mm at isocenter, and the source-to-isocenter distance is 500 mm. The projections are reconstructed with a Hanning filter with a cutoff of 10 lp/cm. Our image values are clamped between -1024 and 15383 HU. A monoenergetic simulation was also done for comparison, with a nominal CT value of 30,000 HU (μ corresponding to 80keV).

While simulations of metal artifacts have been undertaken before [8], the shape-dependence of the artifacts was not investigated, nor was accuracy of reprojections through metal. Our simulations, discussed in this section, explain the appearance of metal artifacts and show how reprojection through the metal cannot

quantify the metal.

Fig. 1.6 shows reconstructed images with and without beam hardening. When the object cross-section is circular, artifact does not exist outside the object. However, beam hardening is visible within the object in the form of cupping. As the object's eccentricity increases, the artifact amplitude outside the object increases. The dark artifact is along the long axis of the ellipse, and the bright artifact is along the short axis.

We have not simulated scatter but make a brief note about it. A scatter event occurs when a primary photon interacts with a particle; a secondary photon is emitted with a lower energy in a new direction and falls onto a different detector element than the one intersecting the original ray. To integrating detectors (without energy discrimination), this event cannot be distinguished from a primary photon detection along the original ray. The apparent attenuation along that path is decreased because of the false event. CT scanners usually have anti-scatter plates, fins of tungsten, that focus to the x-ray source. The anti-scatter plates reduce most of the scatter. Scatter works in the same direction as beam hardening, i.e., the measured attenuation is lower than the ideal attenuation. Scatter simulations are usually done with Monte-Carlo simulations.

Fig. 1.7 illustrates how the artifact arises from inconsistency in the projections. Ideal (monoenergetic) and hardened (polyenergetic) projection data filtered with the Ram-Lak kernel are shown next to the CT image. With a greater ray length through the hardening material, there is a bigger discrepancy between hardened and ideal projections. In the ideal case, the undershoots of the filtered projection data in all angles perfectly compensate for the data backprojected across the image. Consider the central ray in the hardened case. Its amplitude drops less relative to the ideal in the shorter path (0 degrees) than in the longer path (90 degrees). The undershoots are even smaller along the longer paths relative to the ideal case (due to more hardening) than the shorter paths. This leaves the shorter paths with relatively less negative compensation, leading to the bright artifact, and the longer paths with relatively too much compensation, leading to the dark artifact.

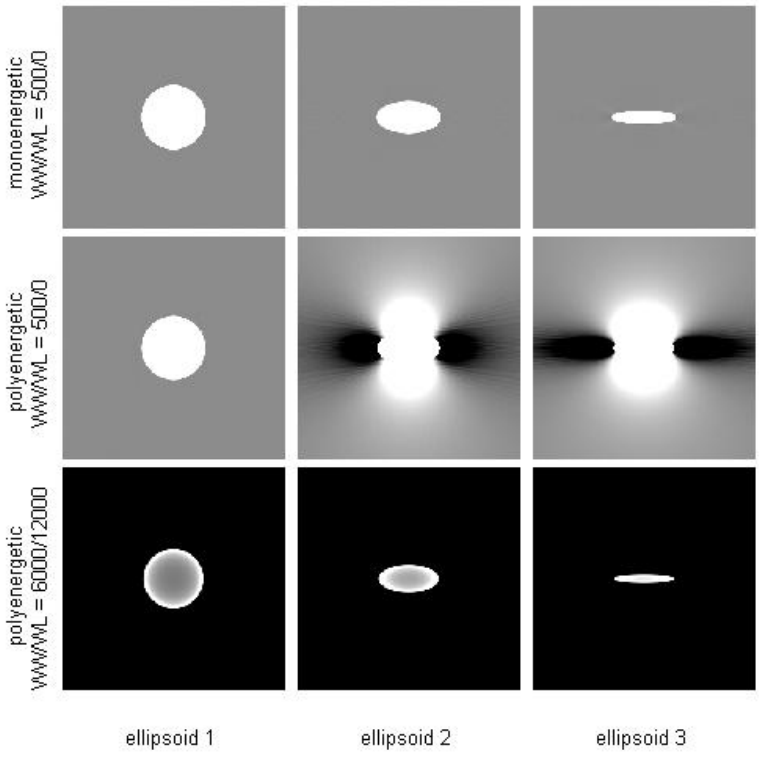


Figure 1.6: Illustration of how beam hardening artifacts change with object shape. The top row shows an ideal (monoenergetic) simulation, and the bottom two rows show simulated beam hardening. For round objects, beam hardening is visible only within the object. This is the well-known cupping artifact. As the metal object eccentricity increases, the artifacts increase as well and are visible outside the object. The dark artifacts are along the long axis (maximum projections). Top two rows: Window Width (WW) / Window Level (WL) = 500 / 0, bottom row: WW / WL= 6000 / 11000 HU.

Note that the attenuation by metal cannot be quantified by simply reprojecting the metal voxels in a reconstructed image. The hardening of the beam causes overestimation of the reprojections in the direction of highest attenuation, and underestimation in the direction of lowest attenuation. Fig. 1.8 shows that the discrepancy between original hardened projections and reprojections increases with eccentricity. This is because the metal object voxels are reconstructed using projections with different amounts of hardening in each view, but the voxels summed in each reprojected view are the same voxels. Considering the central ray along the long axis, the voxels along it were reconstructed using relatively less-hardened

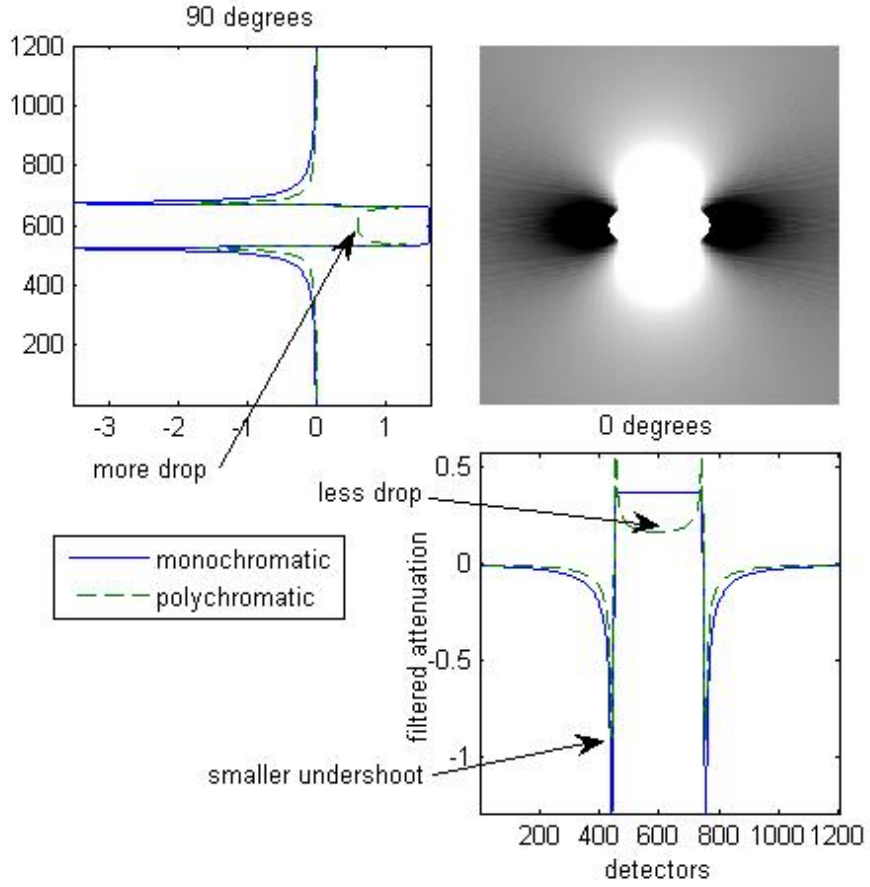


Figure 1.7: Formation of artifacts by beam hardening.

projections along with the more-hardened projections. When compared with the most hardened projections, the reprojections will therefore be greater. A similar reasoning exists for every other ray through the metal. Due to measurement error, a simple beam hardening inversion cannot be performed as is done for water, or for bone correction in head images in which the skull is roughly ring-shaped in each axial slice [11].

It is sometimes thought that noise due to photon starvation is the dominant cause of metal artifacts. However, images from modern CT scanners are not usually limited by photon starvation. These scanners have more powerful X-ray tubes and use a variety of modulation techniques to obtain enough power (photon counts) to reduce noise while limiting dose. In addition, adaptive filtering is applied in

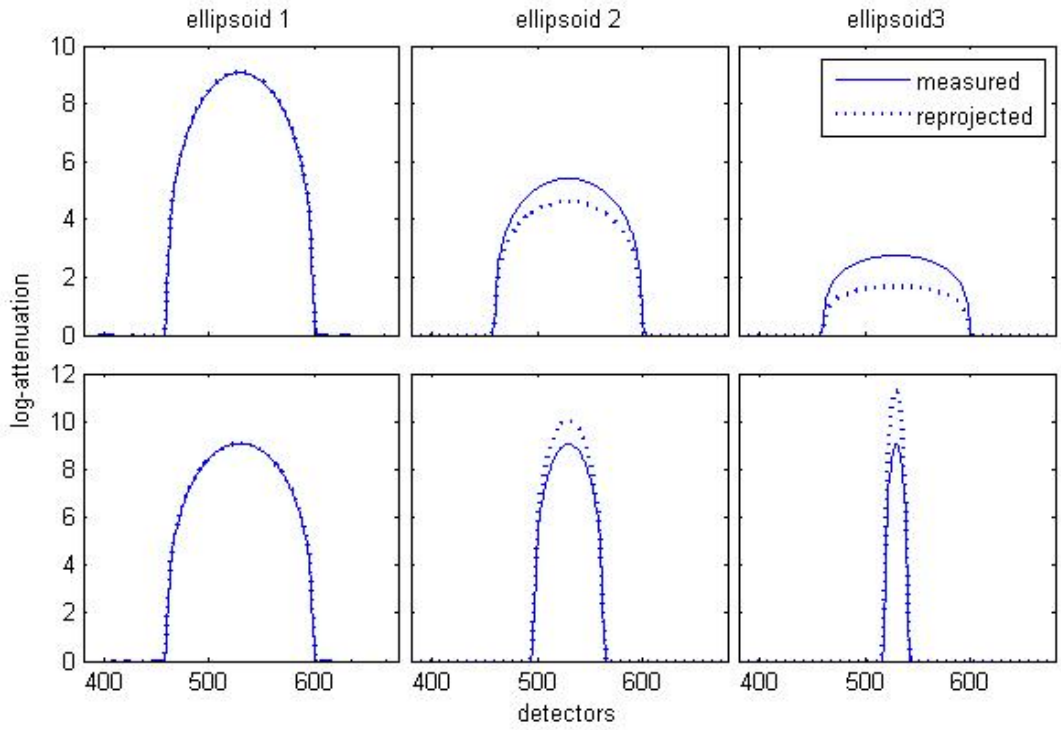


Figure 1.8: Reprojections of the thresholded ellipsoids shown in Fig. 1.6(dotted lines) along with the original hardened projections (solid lines). The top row shows reprojections along the short axis of the ellipsoids, and the bottom along the long axis. Reprojections are underestimated along the short axis and overestimated along the long axis of the ellipsoid.

the scanners to reduce noise to acceptable levels [15]. In our images in Chapters 2 and 3, we do not encounter photon starvation.

1.2.3 Literature review

MAR algorithms often replace the inconsistent projection data, i.e., data from rays passing through metals, with estimates of the true underlying projection data, but when these data estimates are inaccurate, secondary artifacts are generated. The secondary artifacts may be as unacceptable as the original metal artifacts; therefore, accurate data estimation is critical.

MAR algorithms have been developed in medical CT imaging since the 1980s [12]. Despite the advances, there is no widely accepted solution, and MAR

continues to be a challenging research problem. There are three main approaches to MAR algorithms - sinogram completion [12–21], multiple-energy decomposition [22–28], and iterative reconstruction [23, 29–34]. All these methods operate in Radon space, also called the sinogram, or called projections.

Sinogram completion has been the most widely explored because of its low complexity [35, 36]. In sinogram completion methods, the data samples collecting x-rays attenuated by metal are identified in the projection data. These samples are called metal traces. In some methods [12, 13, 15, 16, 27, 37–39], metal objects are located in the original image by thresholding or other segmentation, and the traces are located by calculation. In other methods, [40], the metal traces are located by segmenting projections. Metal trace data are replaced with an estimate of underlying data. The corrected data is used to reconstruct the final image by FBP. These methods are faster than the other approaches, but accurate data estimation within the metal traces is a difficult problem.

Early work [12, 13] interpolated the projection data on either side of the traces. This approach is often called LI-MAR in the literature, where LI is the abbreviation for linear interpolation. We continue to use this name, and apply it to all methods that merely interpolate the data, even when the interpolant may be higher order, or a spline. LI-MAR deletes edges across high-contrast structures, and thus brings about a new inconsistency in projections, leading to secondary artifacts.

Fig. 1.9 illustrates the problem. In the diagram, a background region contains a piece of metal and a high-contrast structure, such as bone. Projections at two angles are shown. In one angle, the bone does not interfere with the metal projection, and the projection can be interpolated without loss of edge information. In the other angle, interpolation of the data would result in an edge being blurred away. Edges missed in projections result in secondary artifacts in images, which may be as severe as the metal artifacts.

It was proposed that edges be recorded from the original image [15]. The edges could be reprojected and subtracted from the scanner projections to create a smoother projection on which interpolation could be performed. However, a

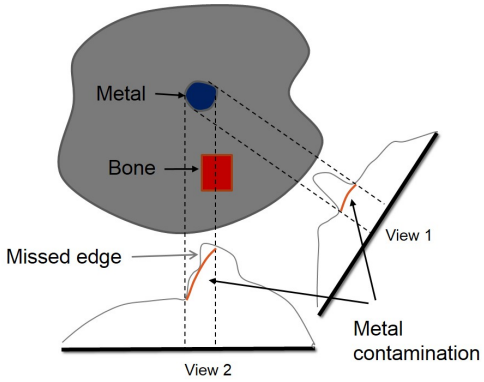


Figure 1.9: Interpolation across the metal traces results in edges being missed.

method for recording edge information in reprojections was not defined, but is critical in the performance of a MAR algorithm. In recent years, in medical imaging, image segmentation has been used to identify high-contrast structures, in order to develop an intermediate image that is often called a prior-image [16,19,21,41]. We hyphenate the words to avoid confusion with true Bayesian priors or the notion of a "previous" image. The prior-image is forward-projected (interchangeably referred to as reprojection) and thus used to guide the data replacement in the scanner sinogram. The segmentation of separate real data from artifacts is a challenging task because the CT density ranges of artifacts and materials overlap, as do their gradient ranges.

It was proposed that the approximate values of image voxels be estimated using k-means clustering and thresholding [16]. Each voxel would be assigned the mean value of its cluster. The reprojections of this prior-image would be used as replacement data in the metal trace. This method was reported to fail when the artifacts were as bright as bone or as dark as air. (This occurred in most of our test images). Other methods built upon the notion of a prior-image that contains edge information. They also differ in how they combine the reprojections of the prior-image with the scanner data. One method uses the ratio of scanner projections to prior reprojections [38], instead of the difference method of [15]. Other methods require repeated reconstructions, each of which improves the prior-image [19, 37]. However, all these methods rely on intensity thresholding to produce the prior-image, and intensity thresholding leads to voxel misclassification in the prior-image.

Voxel misclassification leads to false edges or missed edges in reprojections and therefore to secondary artifacts. Knowledge of the materials and accurate modeling helps reduce some of this risk [20]. We call the thresholded-prior methods TP-MAR, for example [16,19,37,38]. Some TP-MAR methods suggest doing LI-MAR, and then operating on this corrected image to create a prior-image. However the secondary artifacts from LI-MAR can be intense enough that thresholding still leads to voxel misclassification. In a recent work [42], after the publication of our research article), an LI-MAR image is blended with an uncorrected image to create a prior-image, which is shown to give better images in early experimentation.

In segmentation, assumptions are made about the intensity or spatial distribution of human tissue and artifacts. However, no such assumptions can be made about luggage, therefore the medical MAR methods do not work on luggage scans. In the luggage screening application, as we will see, there are enough metal traces that the LI-MAR image is not useful. Non-medical methods are described in [36, 43], that do not construct prior-images, and create less severe secondary artifacts than LI-MAR. Both methods would have better preservation of contrast and artifact suppression with prior-images.

Multiple-energy decomposition methods are used to decompose materials into basis materials. They can compensate for artifacts from beam hardening. Two or more x-ray spectra are required for energy decomposition, and multienergy imaging is not standard in clinical or luggage scanning protocols. Iterative reconstruction of dual energy data has the potential to provide excellent images if dual spectra and models are available [25]. In a different approach, a radiotherapy treatment scan is used to generate a prior-image for use with MAR on a diagnostic scan [44].

Statistical reconstruction algorithms based on scanner modeling have the potential to reduce metal artifacts [23, 29–34]. These methods aim to remove all image degradations through accurate modeling of the x-ray generation, attenuation and detection processes, which is difficult to accomplish, and may even require that the scanned objects be known in materials, shape or both [31,33,45,46]. The assumptions are justified in medical imaging. Their main drawback is that they

are slow.

A recent approach to MAR is to use numerical optimization for reconstruction without detailed scanner modeling, even without noise modeling. This approach applies linear approximations to a non-linear problem. It assumes that the projection data are adequately preprocessed to compensate for other image degradations, but are still degraded by metal. Numerical optimization has become more reliable and efficient in recent years, but its application to MAR is limited [20, 47, 48]. The approach in [47] is formulated along the lines of a previous algorithm to reconstruct images from incomplete projection data by using algebraic reconstruction technique (ART) alternated with steepest descent of the regularization, and by enforcing non-negativity at voxel updates [49]. All the metal trace samples are discarded in [47] to create incomplete data. The images shown in [47] are of circular phantoms with single pieces of metal, and the performance with medical or luggage scans has not been demonstrated.

Like the model-based statistical algorithms, numerical solutions do not preserve texture and desired resolution, depending on the objective function chosen. To correct this problem, in [20] the approach is a hybrid algorithm combining optimization and sinogram completion. They minimize an unconstrained least-squares (LS) objective function regularized by the total variation norm. The minimization is done with an interior-point algorithm [50, 51]. Here again, all metal trace samples are discarded. Since numerical optimization removes texture, the optimum solution is reprojected, and metal traces from the original sinogram are replaced with the reprojected traces. In luggage data, a third or even half of the projection samples may be contaminated by metal. If all these data are discarded, the reconstructions are poor, as we will demonstrate. A second method that estimates and corrects for beam hardening from low atomic number metal is described in [20] but it is cautioned that it is not suitable when partial volume is present. In luggage scanning, we have partial volume in nearly every image, and combinations of metals. A uniformity constraint around metal is imposed in [48], but has limited applicability.

1.3 Evaluation of segmentation algorithms

The U.S. Department of Homeland Security has identified increasing threat categories and lowering false alarms [1] as requirements for future EDS. The segmentation algorithms in the ATR are continuously evolving to meet lower false alarm requirements, and new threat categories. These segmentation algorithms must themselves be evaluated. Incorrect segmentation leads to the apparent merging of different objects, the apparent splitting of single objects, or both.

Quantitative evaluation of segmentation algorithms is a challenging task in luggage screening because multiple splits and merges are possible. In addition to an accuracy score, we would like to gain a deeper understanding of the algorithms' behavior. First, we would like to know if an algorithm systematically oversegments or undersegments images or if the error is random. A knowledge of systematic errors allows us to tune the parameters of a segmentation algorithm, or supplement the segmentation algorithm with additional steps such as region merging [52]. Second, the ability of a segmentation algorithm to capture object features must be evaluated. This is because evaluation of features is critical in ATR. Third, since it is often more important to correctly segment some objects than others, a method to assign priorities to segments is desirable when evaluating the algorithm, for example based on image intensity, homogeneity, particular texture or any other image features that define objects of interest. Fourth, a segmentation algorithm may have varying accuracy across the feature range, and this knowledge can be used to establish confidence in a given segment. There can be no restriction on the number or nature of objects. All these considerations are important in luggage scanning but are not addressed by existing evaluation methods. A review of the existing methods is in Chapter 5.

To encourage the development of new segmentation algorithms for CT security systems, a database of CT images of suitcases was generated by the ALERT group at Northeastern University, and distributed to five research groups at universities and corporations [53]. The database contained no threats; the requirement was to segment all objects present in each suitcase. Segmentation results for a sample of this data were obtained for detailed quantitative evaluation. In this

application, objects missed by the segmentation algorithm correspond to type II error (false negative) in binary classification and spurious objects created by the segmentation algorithm correspond to type I error (false positive). Our evaluation methods were tested on these machine segmentations.

Chapter 2

MAR in medical imaging

As discussed in Section 1.2.3, most MAR methods are based on sinogram completion. These methods estimate projection data within the metal traces, and the accuracy of data estimates determines reconstructed image quality. The data estimates should capture the projection of high contrast edges, otherwise secondary artifacts result from inconsistencies with other data samples. Therefore, recent methods construct an estimate of the image underlying the artifacts by thresholding the original image or an LI-MAR image (i.e, no prior-image). The image estimate should capture high contrast details of the original image. This image estimate is called a prior-image, which is forward projected to obtain projection data estimates. The data estimates in the trace are combined with scanner data in the rest of the sinogram. These methods are vulnerable to the misrepresentation of anatomical edges that reproject onto metal traces.

We have focused our work on the construction of a more accurate prior-image than defined by previous literature. As we will show later, the prior-image has a bigger impact on the final image than the data combination method, assuming a reasonable combination method is used.

2.1 Methods

Our method to build a prior-image operates on the original image. The goal is to discriminate artifacts from real structures in the original image so that we

can replace artifact-contaminated regions of the original image with tissue values, thus generating a prior-image. The separation of artifact from real tissue is not a trivial task. The CT number (voxel intensity) ranges of metal artifacts and anatomical structures overlap, as do their gradient ranges. We base our method on three observations about metal artifacts in CT images, which are supported by the simulations in Section 1.2.2:

1. The artifacts are adjacent to metal pieces.

2. The amplitude of the artifacts decreases as the distance from the metal increases.

3. Local maxima through metal in projection space correspond to dark artifacts in the image.

Based on observation 1, we extract local image extrema that are adjacent to metal voxels and smaller than a given scale, and replace them with values from surrounding voxels. We then use region growing to group all the replaced voxels into labels. The region growing algorithm decides whether to include a voxel in a label using a threshold that depends on the voxel's distance from metal, which makes use of observation 2. The distance-based threshold limits the inclusion of anatomy into the artifact labels. Also based on observation 2, when a local maximum label contains both artifact and bone voxels, we use a discriminant curve to classify voxels within it as belonging either to artifact or to bone. We restore the bone voxels to the image. Based on observation 3, we locate local maxima through metal in projection space, and match them to local image minima. As a result, we can interpret image minima as artifacts even when they are split off from the metal. The dark artifacts can split off from the metal in the presence of high density CT objects, or multiple pieces of metal. The approach, implemented in MATLAB Version 7.10 (The MathWorks Inc., Natick, MA), has the following steps. Fig. 2.1 shows outputs from each step.

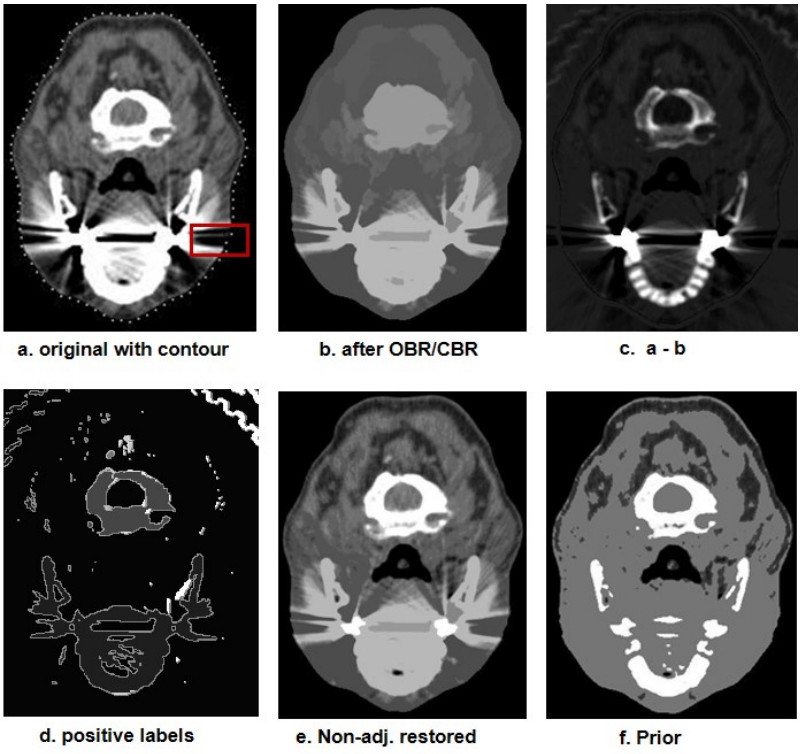


Figure 2.1: Illustration of the method to generate the prior-image of a CT head scan. The contour is represented by the dotted line in (a), the image after CBR and OBR is in (b), the difference image (ab) is in (c), labels from the positive differences are in (d) the recovery of non-adjacent labels is shown in (e) and the final prior image is in (f). WW/WL = 200 / 0 HU in (a), (b), (e) and (f). WW/WL 200/ -1000 HU in (c).

Step 1: Segmentation of metal

Metal voxels are segmented by region growing. Seeds for region growing are voxels above 7000 Hounsfield units (HU). Neighboring voxels are successively added to the region if they are above 3000 HU. Teeth, which have the highest CT intensities for human tissue, are usually less than 3000 HU, so that anatomy is not included in metal labels. Labels are generated for each connected metal region. We then contour the anatomy to prevent dark artifacts, if present, from blending into the surrounding air. Fig. 2.1(a) shows the contour by a dotted line. If dark artifacts blended into the surrounding air (as shown in the rectangle), they would not be interpreted as local minima. The contouring method we used is described

in Appendix A.1.

Step 2: Removal of local maxima and minima

Metal artifacts create local maxima and minima around the metal. Local maxima and minima are respectively removed using closing-by-reconstruction (CBR) followed by opening-by-reconstruction (OBR) [54]. CBR is a morphological operation that performs grayscale dilation with a structuring element followed by iterative erosion that is constrained by the original image. Similarly, OBR first performs grayscale erosion followed by iterations of dilation constrained by the original image. CBR eliminates dark regions (local image minima), and OBR eliminates bright regions (local image maxima) that are smaller than the scale determined by the structuring element. Larger extrema, or regions without extrema, are left alone. CBR and OBR respectively replace voxel values in the closed or opened regions with values derived from voxels surrounding these regions. The structuring element should be at least twice as large as any metal piece in the image to use replacement values outside the artifacts. In order to prevent a too-large structuring element from flooding a local minimum with diffuse bright artifact values, we recommend clamping the image at a high soft tissue value (100 HU) and restoring the image values after CBR. The OBR and CBR operations will also remove anatomical structures that are smaller than the scale of the structuring element. Fig. 2.1(b) shows the result of OBR and CBR, where anatomical structures have been eliminated along with artifacts. We restore the anatomical structures to the prior-image by using the following steps to discriminate between anatomy and artifacts.

Step 3: Recovery of non-adjacent anatomical structures

The OBR and CBR processed image is subtracted from the original image. In this difference image (shown in Fig. 2.1(c)), small intensity differences, attributed to noise or artifact, are eliminated by thresholding. We used a threshold of three times the image noise, which was about 20 HU for our head images. Next, the positive and negative differences are considered separately. Region grow-

ing is performed on the negative voxels of the difference image, using an inter-voxel intensity threshold that depends on distance from the metal. We use a distance-dependent threshold because from observation 2, we can expect that inter-voxel variations in artifacts decrease as distance from the metal increases. The distance-dependent threshold limits the grouping of artifacts with anatomy. We generate a distance transform of the image, which is the smallest distance from each voxel to any metal voxel. Our distance-dependent threshold is defined as a function of the distance transform, as

$$T_\delta(x) = \max\left(Te^{-aD(x)}, T_{min}\right), \tag{2.1}$$

where $D(x)$ is the value of the distance transform at location x, and T, T_{min} and a are constants. We choose $T = 5000$ HU to accommodate the large variations in or near the metal, $a = 0.05$, and $T_{min} = 100$. All values were determined empirically, but are not critical as shown by experiments (varying a from 0 to 0.2) and T_{min} between 50 and 200 HU. Similarly, region-growing is performed on the positive differences. We use the same equation and parameters for region growing of the positive voxels. From region growing, we get labels of positive or negative polarity. For example, positive polarity labels are shown in Fig. 2.1(d). If the labeled regions are not in the neighborhood of a metal label, they are interpreted as anatomical structures, and the voxel values of the original image are restored in those regions. Fig. 2.1(e) shows the recovery of labeled regions that are not neighboring the metals. The artifact labels may border metal pieces or be separated by interference patterns. We define a neighborhood size of 10 mm.

Step 4: Recovery of adjacent anatomical structures

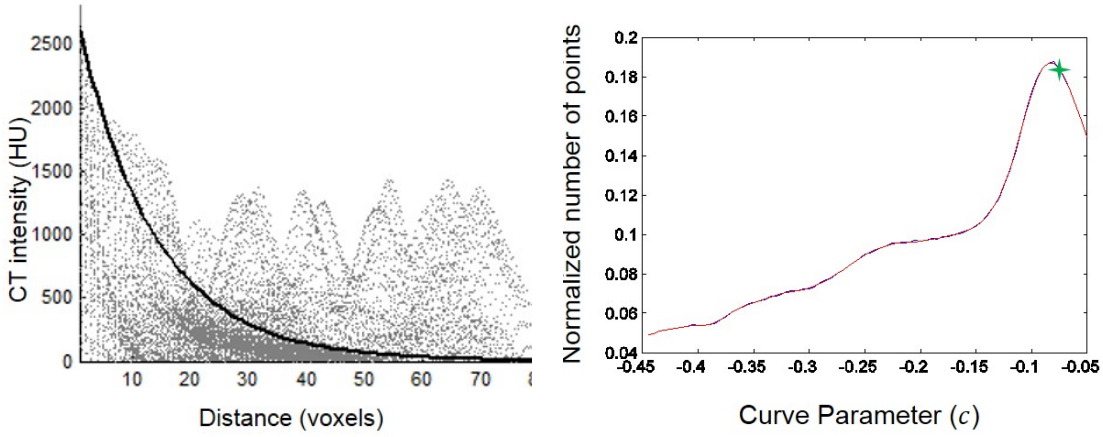
If a region of positive artifact grows into bone, voxels containing bone would be included in the labeled region and incorrectly replaced with soft tissue values. We exploit observation 2, i.e., that artifact amplitude drops as a function of distance from metal. In the positive-polarity labeled regions that remain after the recovery of non-adjacent labels, voxel values of the original image are plotted against the distance transform values. Fig. 2.2(a) shows an example plot. The artifacts

generate a cluster in the plot. To separate the artifact cluster from non-artifact voxels, a set of exponential curves is generated with different parameters. The equation for the family of curves is

$$I_c(D) = (I_{max} - I_{min})e^{-cD} + I_{min}, \tag{2.2}$$

where c is the curve parameter, I_{max} is the maximum value in the region, and D is distance (distance-transform value). I_{min} is the minimum of the region values and an outlier bound of 200 HU. The outlier bound value was chosen because the minimum CT number within the artifact cluster was always about 200 HU in our test cases. In any case, lower artifact intensities are removed in a subsequent step. We used no outlier bounds for the upper CT value.

For each curve, the number of voxels under that curve is normalized by the area under the curve. The normalized number of voxels drops past the curve that includes the cluster as shown in Fig. 2.2(b). We choose a curve that is past the peak, to ensure we have captured the cluster, even at the expense of some anatomy. Voxels above this curve are recovered because they are more influenced by anatomy than artifact.



(a) (b)

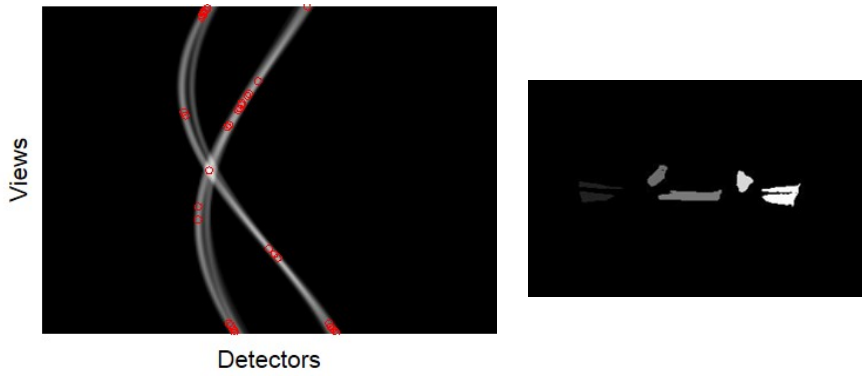
Figure 2.2: Clustering to separate artifacts from anatomy. The relationship between the image intensity and the distance transform for the image in Fig. 2.1 is shown in (a). The line shows the exponential curve with the best separation of artifact from anatomy. The number of points under each exponential curve, normalized by the area under the curve, is shown in (b).

In order to determine whether artifact has indeed grown into anatomy, we compute, from the intensity distance plot, the highest CT value at each distance. We use distance bins of 1 mm. Then we compute the skewness of the resulting distribution. If the skewness is less than 0.15, we assumed that these labels must have grown into bone. The skewness threshold was selected because in our images, the skewness values were about 0.3 in images where the artifact did not grow into bone, and from -0.1 to 0.1 when it did.

Step 5: Deletion of dark artifacts

From observation 3, we can identify local maxima in the projection data and match them with minima labels (i.e., local minima in images). We compute centroids and eigenvectors of minima labels. For each of these labels, we reproject its centroid in the direction of its largest eigenvector. If the centroid projects onto a point that is in the neighborhood of the local maximum of the metal trace, then the label is considered an artifact. We have used a neighborhood size of 10 degrees and 5 mm. To determine the local maxima in projection space, we reproject only the metal. Fig. 2.3 shows the metal traces. In each projection view, we find the maxima in the sample direction. There are one or more maxima in each view. For each trace and for each view, we extract the local maxima values. We fit a sliding polynomial to the local maxima values extracted at each view. Then we locate the maxima of the fitted curve. The sliding polynomial reduces spurious maxima which may appear due to noise or sampling errors. We do not use a one-to-one correspondence of each local-minimum label with a local maximum in projections as a requirement for classifying that label as an artifact, because with multiple projection maxima, negative artifact regions may be combined by region growing. However, the step allows us to add the minima labels that have broken off from the metal. We cannot apply this rule for the bright artifacts (image local maxima), because they do not correspond to local maxima in projection space, rather, they correspond to regions between the local maxima, and are more diffuse in appearance than the dark artifacts. The processed image created thus far is thresholded by the method in Appendix A.2. The segmented metal voxels

are restored, because each metal piece is a real structure, not an artifact. This completes the generation of the prior-image, shown in Fig. 2.1(f). This prior-image is reprojected, and the reprojections are used in interpolation, as described in Appendix A.3.



(a) (b)

Figure 2.3: Reprojections through metal voxels. The local maxima are shown by red circles in (a). The image minima isolated by OBR and CBR are shown in (b).

2.2 Results

Our method was tested on axial head CT scans. Figs. 2.4 and 2.5 each show four cases. In cases 1, 4, 5, 6 and 7, metal artifacts are produced by metal coils in cerebral aneurysms. In case 2, metal artifacts are produced by a deep brain stimulator, and in cases 3 and 8, by dental fillings. The original images are shown in the top row, and images corrected by our MAR algorithm are shown in the second row. For comparison, results are also shown for LI-MAR and an exemplified TP-MAR. This LI-MAR is an improvement upon the original algorithm that uses linear interpolation [13], because this LI-MAR has identical data-fitting and blending to our MAR method. In LI-MAR, since metal traces are deleted, metal pieces are removed in the images. Therefore we have added metal pieces back to the LI-MAR images.

The prior-image for TP-MAR was created by thresholding the LI-MAR image. Although TP-MAR algorithms vary, we used CT thresholds recommended

in [16, 38] for threshold values. Then we used the method for data replacement in the metal traces described in Appendix A.3. Since the data replacement method was the same for all the MAR algorithms, the prior-image determined the improvement.

Artifacts are removed by our algorithm even for multiple metal pieces, and from large metal pieces which produce dark artifacts below -1000 HU (arrow 1), and bright artifacts with the CT intensity of bone or cartilage (arrow 2). There were fewer secondary artifacts with our method than the others. There are residual artifacts in the dental images, especially in Fig. 2.5 (arrow 3), resulting from the imperfect separation of the artifact from teeth. LI-MAR produces secondary artifacts comparable to the original metal artifacts. The TP-MAR images are better than LI-MAR but not as good as those with our prior-image.

2.3 Discussion

2.3.1 Analysis of results

Our algorithm preserves anatomical structures in the prior-image, which is why secondary artifacts are reduced. LI-MAR loses edge information in the metal traces, so estimated data are inconsistent with the rest of the sinogram and secondary artifacts are generated, as shown by arrow 4 in Fig. 2.4. TP-MAR also loses edge information as described later, but to a smaller extent than LI-MAR. Our method operates at a region level while TP-MAR methods operate at a voxel level. A region is more informative than a voxel in distinguishing artifacts from anatomy because the region can be examined against more criteria than just a threshold criterion. We identify a region with constant polarity, and test if the region or a part of it fits the criteria for artifacts. Considering the positive regions, we use the Intensity/Distance relationship (observation 2) to separate bone from bright artifacts. Negative regions can be matched with local maxima in the Radon transform (observation 3) to identify them as artifacts. A TP-MAR prior-image is generated by thresholding an image into air, soft tissue or bone. With misclassification of artifact voxels as bone or air, large errors may be produced in the final

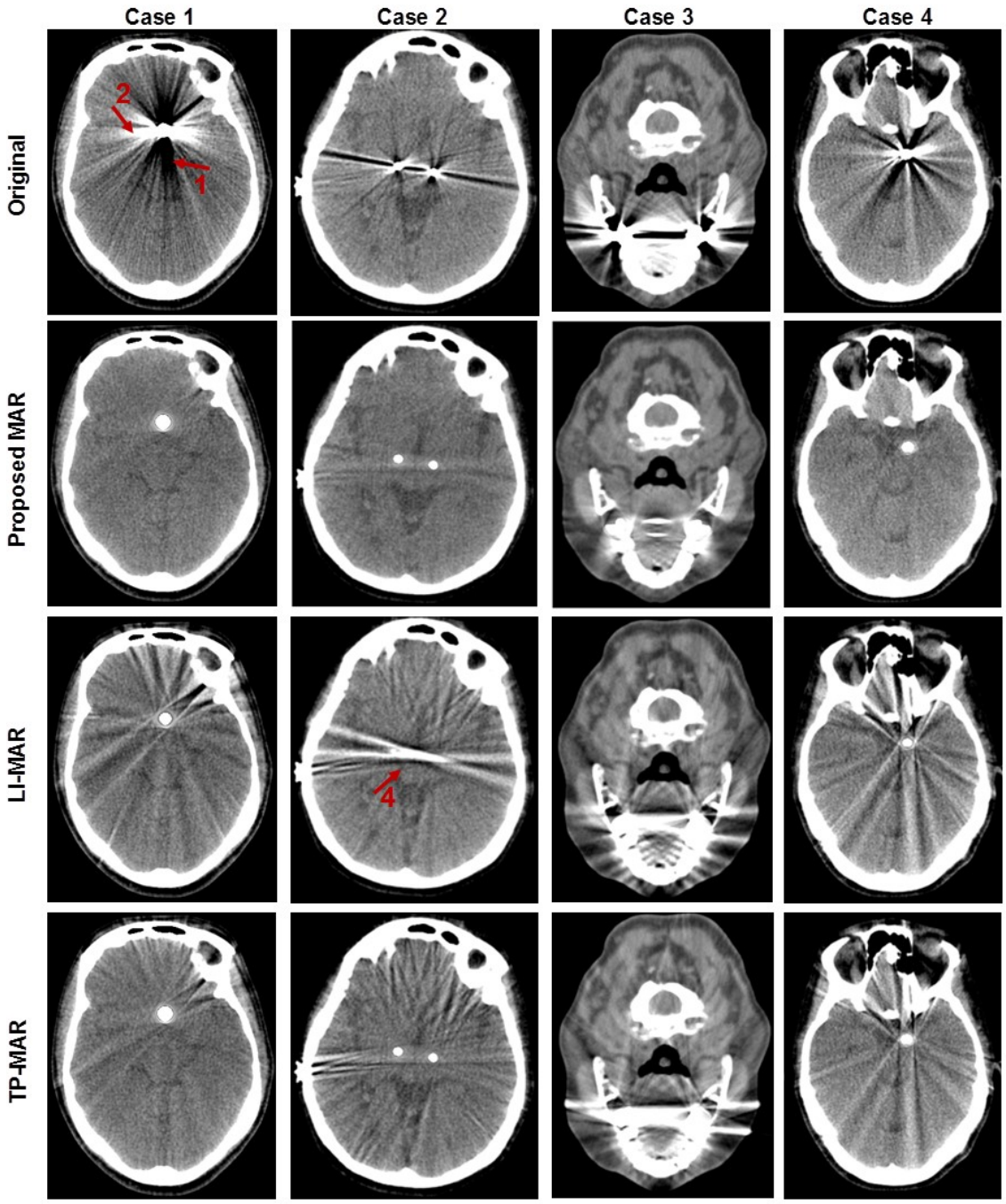


Figure 2.4: Four sets of images are shown in columns. Each set contains the original, the proposed method, LI-MAR and TP-MAR images. In all cases, the proposed prior-image gives best results. WW/WL=200/40 HU for cases 1, 2 and 4, and =500/0 HU for case 3. Arrows point to examples discussed in the text.

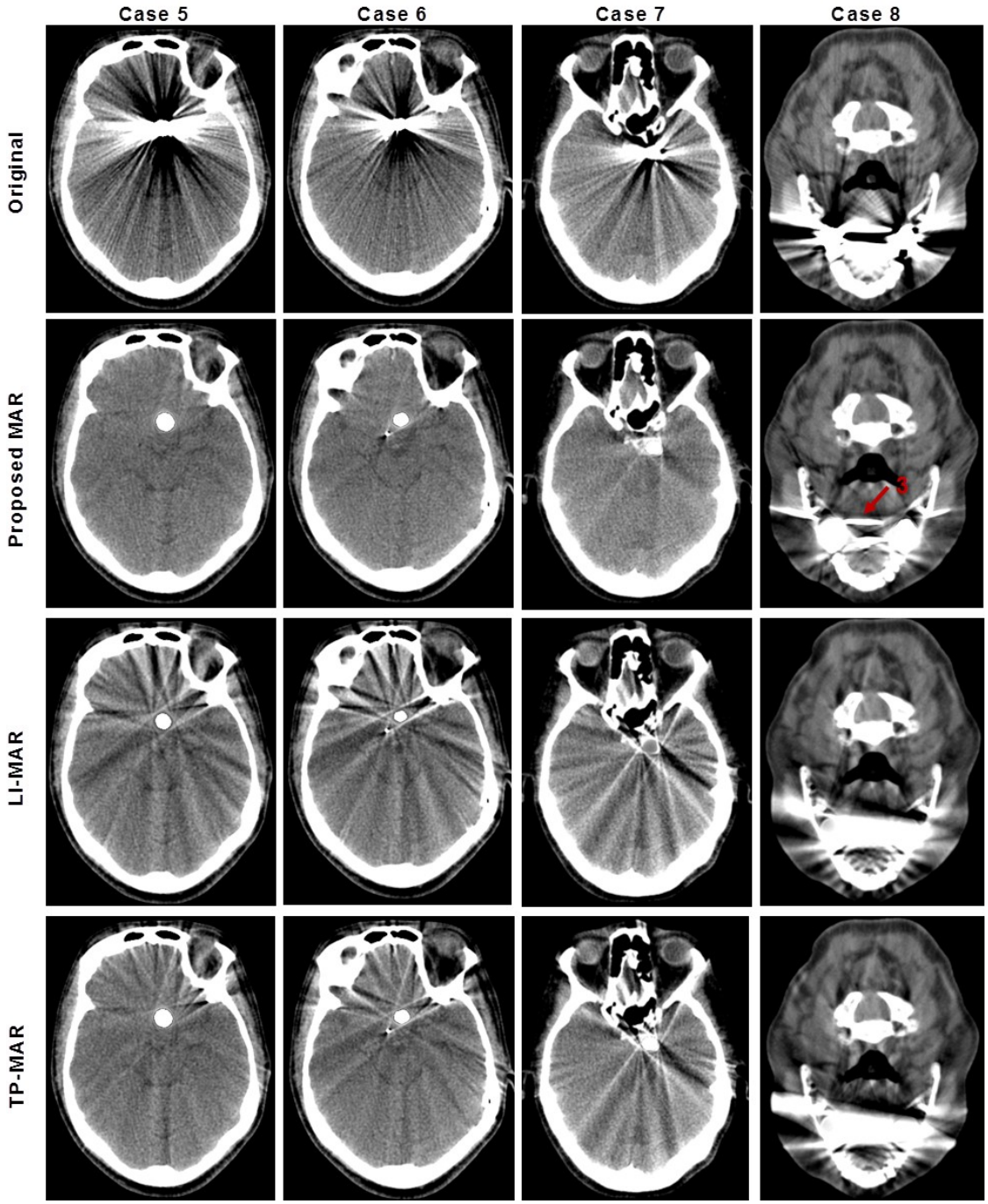


Figure 2.5: Another four sets of original and MAR images. WW/WL = 200 / 40 HU for cases 5-7 and =500 / 40 HU for case 8.

corrected image, especially when the ratio interpolation (discussed below) is used. Therefore, the TP-MAR methods recommend a first-pass LI-MAR to provide an image with smaller amplitude artifacts, so that thresholding of this corrected image is more likely to create a good prior-image. However, our experiments show that the secondary artifacts with LI-MAR may be comparable to or worse than the original metal artifacts. Using poor quality LI-MAR images yields poor prior-images. In the deep-brain stimulator case with multiple metals (Fig. 2.4, case 2), the LI-MAR secondary artifacts are worse than the metal artifacts in the original image (arrow 4). This leads to a prior-image that is worse than the prior from thresholding the original image. In the dental cases (Cases 3 and 8), the secondary artifacts are misclassified as anatomical structures by thresholding, and preserved or enhanced in the final TP-MAR corrected image. Fig. 2.6 shows the different prior-images responsible for the image quality in Case 4. The original image was corrected with LI-MAR. The LI-MAR image was then thresholded to produce a TP-MAR prior-image. Note the partial loss of bone and air pocket structures in the thresholded prior-image (arrow 5), and better preservation in the proposed prior-image. The partial loss is why the image quality of the TP-MAR image is in between that of LI-MAR and our method. The LI-MAR algorithm itself results in the loss of some edge structure, especially if those structures are close to the metal pieces.

In order to study the impact of the interpolation technique relative to the accuracy of the prior-image, we also use the ratio interpolation [38] with our prior-image and with the thresholded prior-image. This interpolation uses the ratio of scanner projections and reprojections, so that the reprojections themselves need not be substituted in the trace. This is a potential improvement on methods that directly use the reprojections in the metal trace, e.g., [16], because reprojections may themselves be inaccurate. The interpolation technique results are shown in Fig. 2.7. Comparing with the difference interpolation (i.e., our data replacement) shown in Fig. 2.4, we see that in the case of a single metal object, the image quality of the difference and ratio interpolation methods is nearly the same when our prior-image was used. The image quality of the ratio interpolation image is

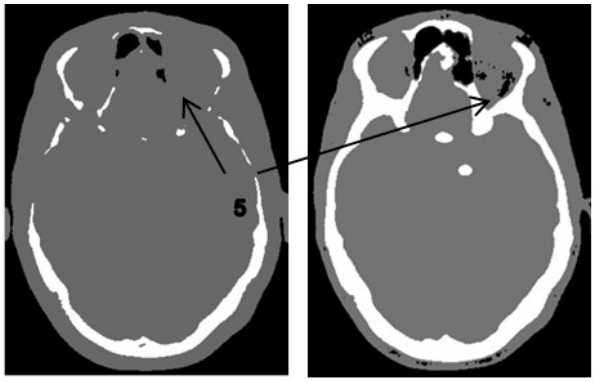


Figure 2.6: Prior-images corresponding to Case 4. The left image shows a TP-MAR prior-image created by thresholding the LI-MAR image Fig. 2.4, row 3. The loss of anatomical edges near the metal leads to a final TP-MAR image that is not much better than LI-MAR in this example. This is because both LI-MAR and TP-MAR have missed nearly the same structures. The right image is the proposed prior-image, and preserves more of the anatomical structure.

nearly the same as the difference interpolation also when the thresholded prior-image was used. These two results indicate that the prior has greater impact than the particular interpolation method. Note that the ratio method, however, only works well when the metal object is removed from the prior-image. The reason is the same one that we discuss in Chapter 1, i.e., that beam hardening coupled with thresholding creates over-estimated or under-estimated reprojections. These reprojections are naturally consistent across view angles, but not once they are multiplied by the ratio of projections, which are not a constant unless the reprojections exactly equal the scanner projections. Therefore, ratio interpolation works well for images with single metal objects, not those containing multiple ones. For multiple metal objects, the metals should not be left out of the prior-image, because while interpolating the trace of one metal piece, we must consider the interfering edges created by the second metal trace. If we leave them out, we miss edges and create secondary artifacts (arrow 6 in Fig. 2.7). Therefore, the interpolation method itself has an impact when there are multiple pieces of metal.

Metal objects that are close together will have traces that overlap in more views than objects that are far apart. For this reason, we recommend that the metal thresholds be lowered, or the metal objects dilated, so that high frequency

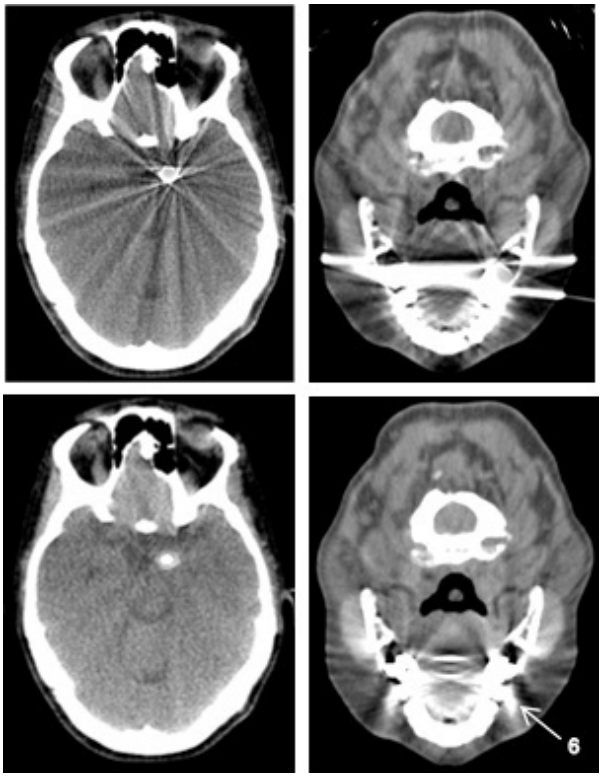


Figure 2.7: Images reconstructed with ratio interpolation [38] for cases 4 (left) and 3 (right). The prior-images for the top row images were thresholded LI-MAR images (Fig. 2.4, row 3), and the prior images for the bottom row images were from our method. Along with Fig. 2.4, these images suggest that the prior-image is more critical than the interpolation technique. However, for multiple pieces of metal, the interpolation technique itself begins to play a role as seen by the degraded image quality of the right column compared to Fig. 2.4.

secondary artifacts are reduced from one metal trace abruptly entering and leaving another trace. However, for metal objects that are close together, we are likely to get some secondary artifacts because we do not correctly quantify the metal pieces themselves. We saw evidence of this error in Case 2 in Fig. 2.4. There is also a blurry region around each piece of metal resulting from the smoothing of data in the spline fit. All MAR algorithms in the literature exhibit similar blurring. A non-linear correction for reduction of discontinuity may be helpful. In Cases 3 and 8, the separation of bone from artifact using a one-parameter curve was imperfect. This is because interference between the four metal pieces makes the distance-intensity relationship deviate from the exponential form we see with

smaller or fewer pieces. Also due to constructive interference, the artifacts were as bright as the metal edges, and were not excluded by thresholding. For these reasons, the image quality of Case 8 is not as good as the other cases. The image quality from the other methods is poor too, because the metal pieces are large, close together, and embedded in dense anatomical structures.

2.3.2 Limitations of this research and future work

One of the limitations of this research is that we have tested our algorithm only on head images. Optimization and testing of our algorithm for head images was considered to be of high impact due to the clinical importance of head CTs, subtle differences in normal and abnormal findings, and the frequent occurrence of metal artifacts. We have set the parameters at each stage based on CT values of anatomical tissue, air and metal. We have also tested the algorithm steps within ranges of parameter values to reduce the risks of over-fitting. However, the parameters (or the criteria to choose them) may need examination or adjustment for other anatomical regions, and further studies to assess the performance of our algorithm on those regions are needed. Another concern is the robustness of the algorithm in the presence of different implants or more pieces of metal. Our data set was limited. The cases we tested had endovascular coils, deep-brain stimulators and dental fillings. These comprise common metal implants in head CT images. More pieces will lead to more complicated interference patterns. We have seen that when multiple pieces of metal are close together (as in Case 8), the clusters in the intensity-distance relationship are not as well extracted by a single-parameter exponential as they are when there are fewer pieces. Multiple rounds of cluster identification, or a superposition of discrimination curves may be required in such cases. In future work, a higher dimensional hypersurface may be investigated, involving the gradient and other predictors of artifacts, for better discrimination of bright artifacts from bone. Consideration of gradient amplitude and phase might help in removing the residual bright artifact in the prior for the dental image.

In future work, we can add to the prior knowledge to build a more accurate prior image. An anatomical atlas may help distinguish between anatomy and

artifact. An instance of where an atlas may have been useful is Case 8. The residual bright artifact in the soft tissue between the teeth in the prior for the dental image could be identified as unlikely to be anatomical. Machine learning techniques could potentially be useful for classifying the labels as artifacts or anatomy. There is much potential for exploration of computer vision and learning algorithms in developing a prior-image.

2.4 Conclusion

Metal artifacts are predictable because phenomena such as beam hardening that cause them are well understood. We have made some observations about the appearance of metal artifacts and used these to segment the metal artifacts from anatomical structures. We have used the results of the segmentation to build up a prior knowledge image that guides the data replacement step for an effective metal artifact reduction. We have tested this concept of using image segmentation and artifact predictability to discriminate artifacts from anatomy on head images. The priors resulting from our method produce better quality images than LI-MAR or prior-images obtained from thresholding other images. We have found that an accurate prior-image has more impact on final image quality than the choice of a particular interpolation technique.

This chapter contains material from the paper "Segmentation of Artifacts and Anatomy in CT Metal Artifact Reduction," *Medical Physics*, Vol. 39, Issue 10, 2012. This paper was co-authored by Pamela Cosman, Harry Martz and Christoph Wald.

Chapter 3

MAR in CT-based luggage screening

3.1 Introduction

In luggage data, a third or even half of the projection sample rays may pass through metal. These samples are inconsistent with the linear (log-attenuation) assumption of Eq. 1.10. If linear reconstruction methods are applied, a model mismatch is present, which causes metal artifacts to appear in reconstructed images. The transform based reconstruction methods are analytical inversion formulas, so intrinsically, they are rigid. In the medical methods, we used assumptions on image contents to estimate the underlying data in the metal traces, combine the original data outside the metal traces with estimated data, and then reconstruct this combined data by FBP. If the inconsistent data are discarded to once again obtain a linear model, we cannot use transform-based techniques to reconstruct images, because the inversion implicitly inserts zeros in the missing data. Zeros are even more inconsistent with the other projections than the original measurements, and analytical reconstructions from such data would be meaningless.

We turn to numerical solutions for their greater flexibility. An image in vectorized form is represented by x and the scanner sinogram by b. Let the forward-projection model for log-attenuation projections be denoted A. Each cell of the

matrix A, a_{ij}, contains the fraction of the voxel j that goes into the measured data sample i. For noise-free data without residual beam hardening, uncompensated scatter, sampling errors, patient motion, cone or helical motion, the following equation holds:

$$Ax = b. \tag{3.1}$$

We consider x to be the estimate of the model parameters.

With inconsistent data, there is a model mismatch. Other sources of ill-posedness are from noise. In prior research, the inconsistent data were discarded to obtain a linear model. This is outlier rejection. Discarding data, however, raises the condition number of A, and creates a null space. Regularization is used to reduce the secondary artifacts that result [20], and non-negativity constraints are used in [47], as is usually done in statistical reconstruction literature [5,6,49,55,56]. However, in luggage, enough projection data is lost that regularization and non-negativity cannot compensate. Regularization is imposed on the voxel neighborhood, so it does not reduce large low frequency errors without reducing edges. We will demonstrate that discarding all the outliers can lead to poor reconstruction.

Our method is also a hybrid method, because we create a prior-image followed by forward projection and FBP. However, we retain the metal projection data but reduce their weights relative to the non-metal data. Our target prior-image is one that is free of metal artifacts, and has sparse gradients. Our work can be viewed as an extension of the ideas in [20, 47, 49] and [15].

To build our prior-image, we exploit the following observations and facts:

1. When there is greater x-ray attenuation through the metal, there is a greater difference between ideal and measured projection data.

2. Beam hardening and scatter result in lower data measurements than the ideal (monoenergetic equivalent) measurements.

3. The artifacts from beam hardening and scatter have substantial low-frequency content.

As will be explained in detail in Section 3.2, we use the first fact to penalize the projection samples according to the amount of metal. The second point gives us a

constraint. We exploit the third observation by formulating two optimization problems and taking their difference to yield an artifact-only image. This observation allows us to reduce the scale of the problem.

We would like to trade-off some of the inconsistency for a loss of details. We would like a solution that minimizes some objective function $f(x)$, such that the error between observed and estimated projection data is small outside the metal traces, and is some function of the quantity of metal within the traces, as shown in Eq. (3.2).

$$\begin{aligned} &\min_x f(x) && \text{s.t.} \\ &||Ax-b||_L < \delta && i \in \overline{\mathbb{M}} \\ &||Ax-b||_L < \epsilon_i && i \in \mathbb{M} \end{aligned} \tag{3.2}$$

where $\epsilon_i = e(\sum_j A_{ij} x_j^m)$ where x^m denotes a vector containing metal voxels, and is an estimate of the true metal voxel values. The abbreviation "s.t." stands for "subject to".

If ϵ_i were $c_i^T x^m$, and $L = 2$, then Eq. 3.2 would be a set of second order cone programs (SOCP), which is already difficult to solve for large problems. In this form, we do not know if there is a feasible set. In addition, if ϵ_i is not a linear function of x, the solution is unclear.

We try to investigate a simpler approach as in the medical imaging application; our solution (parameter estimates) should trade off inconsistency for details in the metal trace data. We can continue to use the observed data outside the metal traces. When the data are combined, they are inconsistent in a different sense because they capture different levels of detail. However, from all the previous research we have cited, this combined data produces acceptable images.

We use the equivalence between the forms

$$\min_x \ f(x) \ \text{ s.t. } \ ||Cx-d||_2 < \delta \tag{3.3}$$

and

$$\min_x \ ||Cx-d||_2 + \beta f(x) \tag{3.4}$$

for some pair β, δ. So,

$$\begin{aligned} &\min_x f(x) \quad \text{s.t. } ||Ax-b||_L < \delta \equiv \min_x ||Ax-b||_L + \beta_1 f(x) \quad i \in \overline{\mathbb{M}} \\ &\min_x f(x) \quad \text{s.t. } ||Ax-b||_L < \epsilon_i \equiv \min_x ||Ax-b||_L + \beta_i f(x) \quad i \in \mathbb{M} \end{aligned} \tag{3.5}$$

If we sum over the measurements, we get a problem of a weighted least-squares (WLS) form, which is a tractable, convex problem:

$$\min_x \frac{1}{\beta_1} \sum_{i \in \overline{\mathbb{M}}} ||A_i^T x - b_i||_L + \sum_{j \in \mathbb{M}} \frac{1}{\beta_j} ||A_j^T x - b_j||_L + f(x). \tag{3.6}$$

We formulate a constrained optimization problem to obtain a prior-image. As mentioned above, a numerical solution may have different spatial resolution than a desired FBP solution. In Section 3.2.1 we first describe a convex optimization problem to construct a prior-image, neglecting resolution considerations, to explain the objective function and constraint. In Section 3.2.3 we describe the practical implementation of our complete MAR algorithm.

3.2 Methods

3.2.1 Prior-image reconstruction as a solution of a convex optimization problem

Eq. 3.6 is a regularized weighted least squares (WLS) problem. A WLS approach is usually applied to de-emphasize samples based on noise, assuming the noise is additive, e.g., [56]. In our case, however, we de-emphasize the samples, not based on noise, but according to the attenuation through metal. The optimization problem is expressed in the following equation:

$$\begin{aligned} &\min_x \quad (Ax-b)^T W_{X_{Orig}} (Ax-b) + \beta ||x||_{TV} \\ &\text{s.t.} \quad I_P(Ax-b) + \sigma \succeq 0. \end{aligned} \tag{3.7}$$

We first discuss the objective function. $W_{X_{Orig}}$ is a matrix of weights that is computed based on an initial (FBP) reconstruction, $x = X_{Orig}$. $W_{X_{Orig}}$ is diagonal,

and given by the following expression,

$$W_{X_{Orig}} = \mathbf{diag}(w(i)) = \exp(-\lambda \sum_{j=1}^{V} a_{ij} I_1(j)), \tag{3.8}$$

where V is the number of voxels, and λ is an experimentally determined constant, set to 0.2. We reconstructed a single image with a few values of λ and selected the value that gave the best visual result. In Eq. 3.8, I_1 is an indicator function:

$$I_1(j) = \begin{cases} 1 & X_{Orig}(j) > M_1 \\ 0 & \text{otherwise.} \end{cases} \tag{3.9}$$

The threshold M_1 is set to 4000 Modified Hounsfield units (MHU). A voxel above this threshold is interpreted to contain metal or be close to metal. The MHU scale has an offset of 1000 relative to the conventional Hounsfield scale, so that water is 1000 MHU and air is zero. We use MHU rather than HU in this chapter so that "non-negativity" has the correct meaning.

The summation in Eq. 3.8 represents the path length through metal of the ray that gives projection sample i. We choose an exponential weighting function because the attenuation of x-rays is exponential [4], and because exponential weights are monotonic and smooth. Further, a one-parameter function is easy to tune. Although $W_{X_{Orig}}$ depends on the image, the first term in Eq. 3.7 is quadratic because $W_{X_{Orig}}$ is computed only once per image.

The regularization term $||x||_{TV}$ is the total variation norm and β is its strength. The total variation norm is popular in compressive sensing, and it has been used for reconstruction from incomplete data [49, 57]. It rewards sparsity of the gradient. Since this norm is the L_1-norm of a linear operator, it keeps the optimization problem convex. Regularization is needed for stability but it also contributes to the reduction of artifacts. However, the artifact reduction is mainly achieved by the weights and constraint.

Now we discuss the constraint in the second line of Eq. 3.7. The symbol $\succeq$ denotes a vector inequality. This is a linear constraint that has not yet been explored in the MAR literature. It is motivated by the knowledge that the metal artifacts are largely due to beam hardening and scatter. These phenomena are

not additive noise. Both work in the same direction: the measured attenuation is lower than the ideal (monoenergetic equivalent) attenuation, neglecting noise.

Consider the attenuation equations Eqs. 1.9- 1.11. In Eq. 1.11, if we normalize I_0 to 1, then $S(E)$ represents a probability mass function for the incident energy. Consider a homogenous object. The expression

$$-\log \int e^{-\mu(E)l} dE \tag{3.10}$$

is a concave function in the path length l since it is of the -log-sum-exp form [58]. This function includes zero, is non-decreasing and is positive. Therefore, the measured value is always less than the ideal.

We discard the non-negativity constraint of previous MAR literature, because the source of the beam hardening and scatter artifacts is addressed by the new constraint. I_P is a diagonal matrix containing a second indicator function for metal.

$$I_P = \mathbf{diag}(p(i)) = \begin{cases} 1 & \sum_{j=1}^{V} a_{ij} I_2(j) > T \\ 0 & \text{otherwise} \end{cases} \tag{3.11}$$

and

$$I_2(j) = \begin{cases} 1 & X_{Orig}(j) > M_2 \\ 0 & \text{otherwise,} \end{cases} \tag{3.12}$$

where M_2 is set to 8000 MHU. Since we have noisy measurements, we make an allowance for noise in the constraint. The term σ_p is a vector containing the standard deviation (SD) of the noise estimated per sample. The estimate of the noise in each sample is derived by established methods [2, 59, 60]. Note that we have used two different metal thresholds. For weighting, we use $M_1 = 4000$ MHU, and for the constraint we use $M_2 = 8000$ MHU. This is because we apply the constraint only for high atomic number metals such as copper or iron, for which signifcant beam hardening is expected. We have not applied the constraint for aluminum, because the hardening from aluminum is smaller, so the constraint may not be robust enough against partial volume and blurring from higher atomic number metal objects which may result in lower CT values. The threshold T represents reprojection through 20 voxel widths. Setting $T > 0$ reduces the chance

of including voxels with high values due to blurring from adjacent voxels, while still including voxels in the interior of high density metal structures whose values are lowered by beam-hardening.

3.2.2 Practical difficulties and solution

The matrix A has $737,280$ rows and $262,144$ columns (details are in Section 3.2.4), and it is about 1% filled. We would like to shrink the size of the problem by solving for a miniature image. We reduce the image by a factor of four in each dimension, therefore by a factor of 16 altogether. Building miniature images allows us also to reduce the size of our sinogram. We low-pass filter the projections in view and sample directions, and downsample by a factor of four in the view direction and four in the sample direction. This miniaturization allowed us to overcome computational limitations. The total reduction in size reduces reconstruction time by a factor of 16^3. The prior-image should represent the attenuation of objects that are dense enough to cause secondary artifacts, such as water, rubber and plastics. The miniature solution can be upsampled to the same size as the original image X_{Orig} and directly used as the prior-image, but we do not do so because although larger structures are preserved, small structures are degraded. Instead, we try to isolate the artifacts and upsample them, because they do not have the detailed structure we would like to preserve.

In order to isolate artifacts, we consider the equivalence between FBP and least-squares (LS) solutions [3]. The Moore-Penrose (M-P) solution for the problem $Ax = b$ is

$$x^+ = (x_k + x_a)^+ = A^+b = (A^TA)^{-1}A^Tb. \tag{3.13}$$

The term A^Tb is backprojection, and $(A^TA)^{-1}$ is equivalent to the Ram-Lak filter. The M-P solution is also the LS solution.

Therefore, the difference between the LS solution and the constrained WLS solution isolates the artifacts. We must solve for x^+, and the WLS problem and take the difference. Another way of thinking about this is if we consider an "artifact-free" image x_k, and an "artifact-only" image x_a. Our solution should be such that x_k models a consistent data set, and $x_k + x_a$ models the observed

data b:

$$\begin{aligned} \min_{x_k, x_a} \quad & (Ax_k - b)^T W (Ax_k - b) + \beta ||x_k||_{TV} \\ \text{s.t.} \quad & ||A(x_k + x_a) - b||_2 < \delta \\ & I_P(Ax_k - b) + \sigma \succeq 0. \end{aligned} \tag{3.14}$$

Eq. 3.14 is equivalent to

$$\begin{aligned} \min_{x_k, x_a} \quad & (Ax_k - b)^T W (Ax_k - b) + (A(x_k + x_a) - b)^T (A(x_k + x_a) - b) \\ & + \beta ||x_k||_{TV} + \beta_2 ||x_a||_{TV} \\ \text{s.t.} \quad & I_P(Ax_k - b) + \sigma \succeq 0. \end{aligned} \tag{3.15}$$

We must solve for x_k and x_a, and retain only x_a. Since solving for twice the number of variables will take longer, and since the constraints are only for one set of variables, we implement the difference of LS and WLS solutions, because the LS solution is unconstrained and can be solved with a faster solver than the constrained problem.

3.2.3 Practical implementation

The description of the implementation is broken up into three subsections, and illustrated in Fig. 3.1.

Identification of metal in the image and sinogram: An FBP reconstruction of the scanner sinogram gives the original image, X_{Orig}. Matlab's standard functions were used for all FBP reconstructions. We use a simple segmentation technique, region growing, to identify image regions containing metal [54]. If a piece of metal has a mass (calculated as CT density times volume) above a minimum mass threshold, its trace in the sinogram is calculated and will be replaced.

Construction of the prior-image: We forward project the metal voxels to calculate the attenuation from metal and calculate the weights from Eq. 3.8. (Note that we can only approximate the attenuation from metal in this way because beam hardening degrades the reconstruction of the metal itself.)

The quadratic program expressed in Eq. 3.7 is solved using the Mosek software (Mosek ApS, Denmark) [61]. A was generated by forward projection software [62]. We denote the optimal solution X_C^{mini}.

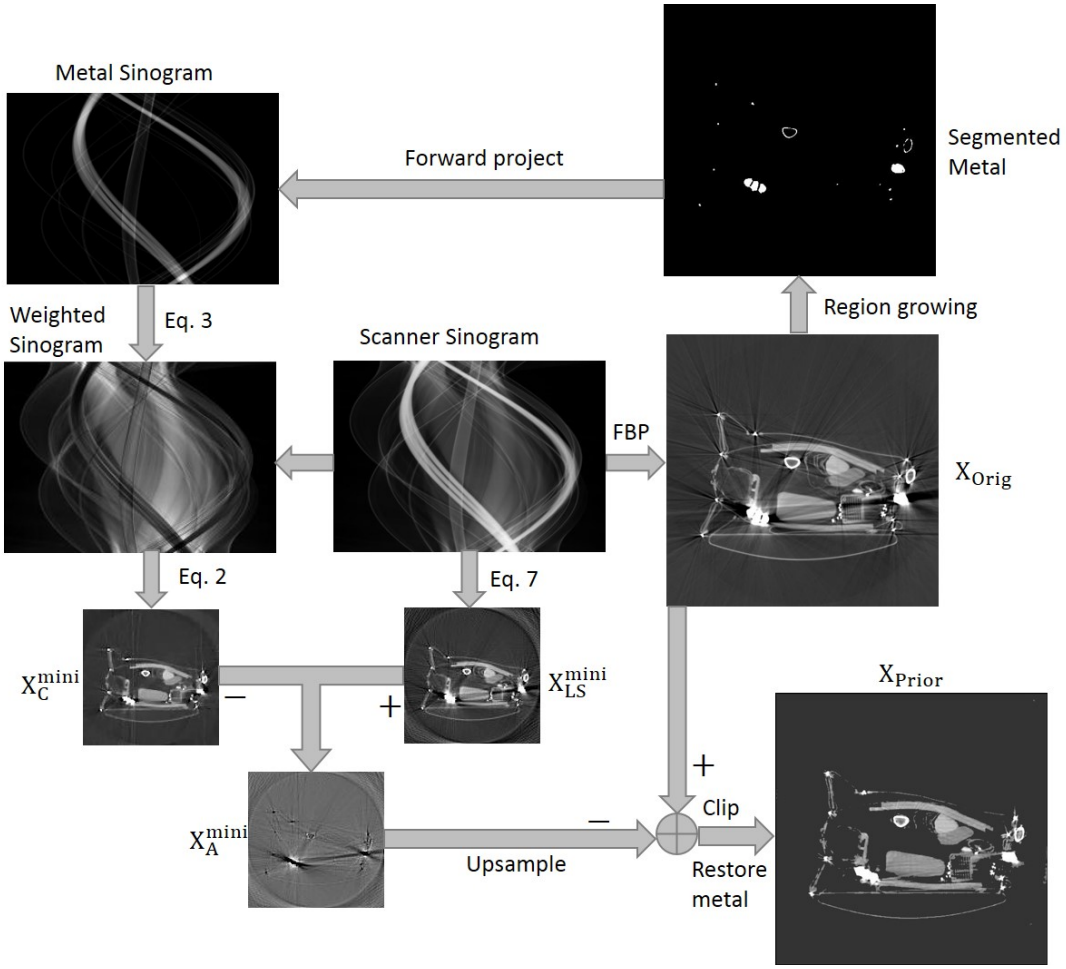


Figure 3.1: Pictorial representation of the construction of the prior-image. The flow starts with the scanner sinogram, shown in the center. The optimal solutions to Eqs. 3.7 and 3.16 are shown in the smaller images (not to scale).

We reconstruct the LS solution, i.e., the M-P solution:

$$\min_x (Ax - b)^T (Ax - b) + \beta_2 ||x||_{TV}. \tag{3.16}$$

Let the optimal solution to the above equation be X_{LS}^{mini}. In Eq. 3.16, there are no weights or constraints. The regularization strength here is $\beta_2 = 0.1\beta$ in Eq. 3.7. We use a smaller strength here because we want less interference with the metal artifact structure.

The difference between X_C^{mini} and X_{LS}^{mini} gives an image consisting mainly of artifacts X_A^{mini}.

$$X_A^{mini} = X_{LS}^{mini} - X_C^{mini} \tag{3.17}$$

We upsample X_A^{mini} using bicubic interpolation to get a full-size artifact only image X_A. Artifacts are removed from X_{Orig} by subtracting X_A:

$$X'_{Prior}(j) = X_{Orig}(j) - X_A(j) \tag{3.18}$$

We used the NESTA solver [51] to solve Eq. 3.16 because it was faster than Mosek. NESTA does not allow constraints. There are two more simple but helpful steps. We copy the segmented metal voxels from the original image to the prior-image. This gives us more accurate trace boundaries. More importantly, metal contributes high contrast structures, which should be preserved in the prior-image. Lastly, we clip the small CT values (i.e., below 500 MHU) to the value of air.

$$X_{Prior}(j) = \begin{cases} X'_{Prior}(j) & X'_{Prior}(j) \geq 500 \\ 0 & \text{otherwise.} \end{cases} \tag{3.19}$$

This removes smaller artifacts and any artificial textures created in low density materials such as clothing. We think of this step as a post-processing operation. There can be residual model mismatch as a result of our trade-off (not discarding all data), and some new secondary artifacts from reducing the contribution of some data. In addition, the discontinuous switching of the constraint could cause some secondary artifacts.

Sinogram completion and final reconstruction: In this step (not shown in Fig. 3.1), we forward project the prior-image, and use a previously published method to replace metal trace data [15]. This method is similar to that in [16] and [21]. In this sinogram completion method, we compute the difference between the original sinogram and the reprojection. We interpolate over the metal trace in these difference projections and get the error. The error is subtracted from the original sinogram, giving us a corrected sinogram. The corrected sinogram is reconstructed with FBP. We have improved upon [15] by fitting the data rather than using linear interpolation, and by blending the corrected with the original sinogram data using Parker weights.

Our original and final image size is 512×512 voxels, and the sinogram size is 1024×720. Due to the downsampling of the sinogram and image, the A matrix size is 46080×16384.

3.2.4 Data and scanner description

Our data set is obtained from the ALERT group at Northeastern University [63] and consists of scanner data from eight bags. Bags were packed with various levels of clutter and included an assortment of metallic objects. There were also metallic bag parts. In each bag, there were some objects with uniform attenuation, e.g., contained liquids. The bags were scanned on an electron beam scanner (Imatron, San Francisco, California). The scan technique parameters were 130 kVp, 63 mAs, axial half-scan, 1.5 mm slice thickness, a 475 mm field of view, 864 fan views and 888 samples per view. The fan views were rebinned to 720 parallel views with 1024 samples/view. There was a 1.3-mm increment during the slice acquisition, and a cone angle of 0.3 degrees. These cannot be compensated with a single-slice half-scan. The scanner projection data were corrected for offset (dark current), gain, and beam hardening by water (water calibration) [64]. The correction was water-based because the Imatron is a medical scanner. Although another calibration material for luggage screening may improve overall image quality, to our knowledge there is no industry standard for a substitute. No scatter correction was applied.

3.2.5 Evaluation

The MAR algorithm was evaluated visually and by three quantitative measures that measure different aspects of quality. The first measure is based on uniform objects. Existing MAR literature has used phantoms with uniform regions to evaluate the effectiveness of MAR [16, 65]. In our case, each bag contains objects with uniform attenuation, such as contained liquids. In the original images, we manually segment those image regions that we know should be uniform. For each such region, in both the original and MAR images, we characterize the CT number distribution in MHU with minimum, maximum, mean and SD. We also measure the Kolmogorov-Smirnoff (KS2) divergence between the original and MAR images.

The second measure is a sinogram-based error [66]. The sinogram-based error is the L2-norm of the difference of original and synthetic sinograms computed

on metal-free samples, normalized by the L2-norm of the original sinogram. The final measure is a reference-free image gradient-based score [66]. The gradient-based score measures the sum of all image gradients. Lower scores correspond to better image quality. We have normalized this score by the value computed on the original images. In addition, we have applied the gradient-based score in 10 pixel-wide bands outside the uniform regions. These bands contain the object boundaries. Since we do not want to lose edges, higher values of this score represent better images.

We evaluate our performance against the iterative projection replacement (IPR) method [20]. We chose this method as a benchmark because of its good results on medical images, and because it makes no assumptions about image content, which makes it more robust than methods specific to medical imaging. We computed optimal solutions of the same size as with our method. This made it possible to reconstruct the images on our 16-processor GPU with 96 GB RAM using their preferred solver, NESTA [51], and optimization problem definition. While the authors do not specify image sizes in [20], they state that resolution matching is not required.

We make some improvements to the IPR prior. We copy the original metal into the prior-image, and set any negative voxels to the value of air (0 MHU). These are trivial changes with large improvements in image quality, and are done in most sinogram completion methods. IPR substitutes the metal traces in the scanner sinogram with reprojected prior-image traces. Substitution may result in discontinuities at the edges of metal traces, hence it may not give good data estimates [21]. Therefore, in addition to IPR results, we compute and present the results of using [15] for sinogram completion, along with the clipping in Eq. 3.19 to further reduce blurring, and call this IPR+.

3.3 Results and discussion

We first show one image reconstructed in various ways, in order to qualitatively explain the image quality improvements from our method. Next, we present

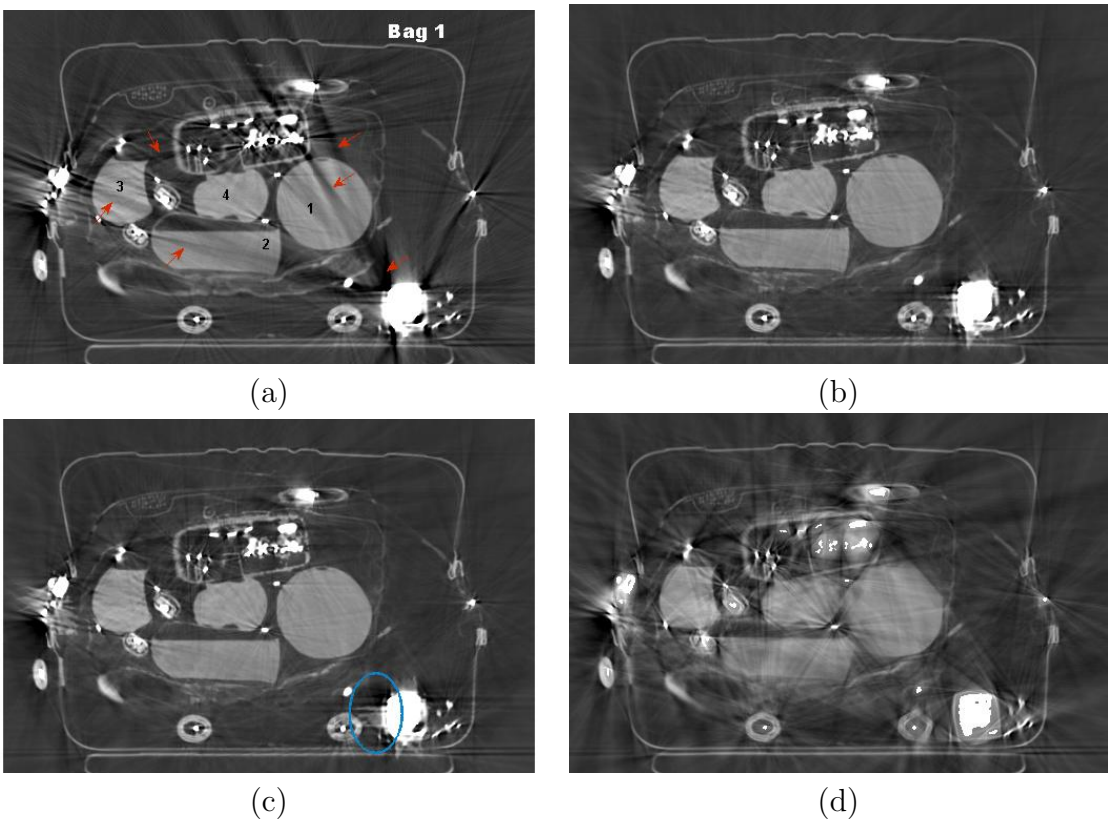
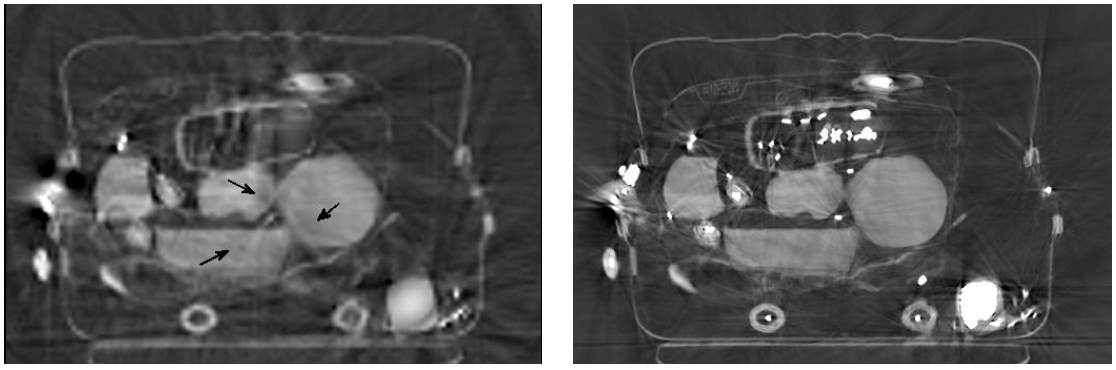


Figure 3.2: Reconstructions of an image from Bag 1. (a) Original image with arrows pointing to the metal artifacts, (b) our method, (c) unconstrained WLS + TV, and (d) LI-MAR. Window width (WW) = 2500, window-level (WL)=750 MHU. Numbered objects are shown in (a), these denote uniform objects which are quantitatively evaluated in Table 3.1.

original and MAR images from more test cases for visual assessment. Third, we show the quantitative results and compare our results with those of a previously published method visually and quantitatively. Finally, we present additional experiments that give a better understanding of the methods.

3.3.1 Qualitative explanation of image quality

Fig. 3.2 shows a 2D image through one bag. The original image with artifacts is in (a) and the MAR image is in (b). The metal artifacts are visually reduced in the MAR image, and this is later confirmed by the quantitative evaluation. For comparison, a regularized WLS image without constraints is shown in Fig. 3.2(c).



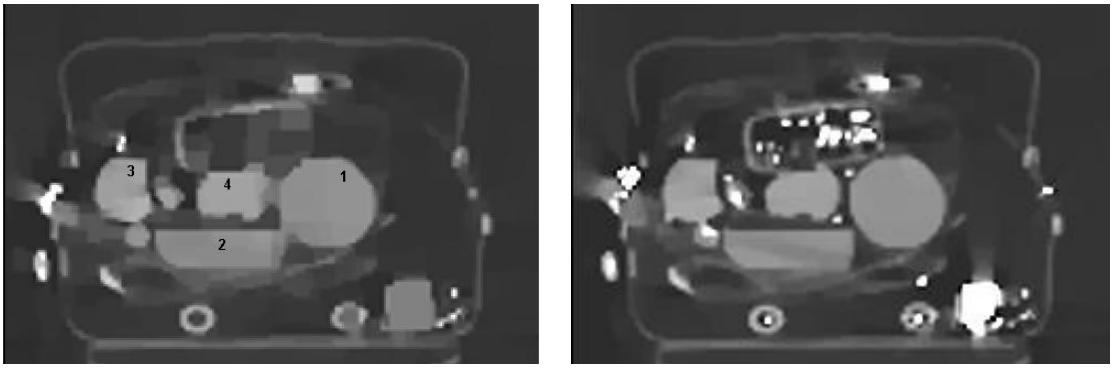
(a) (b)

Figure 3.3: Images showing the effect of discarding all metal-contaminated projection samples. (a) The optimal solution of an unconstrained regularized LS solution. (b) MAR image obtained by using the image in (a) as a prior-image (after metal voxels are restored). WW=2500, WL=750.

This image removes most of the artifacts from the uniform objects. However, the area around the metal continues to show some artifacts, highlighted by the oval. An LI-MAR image is shown in Fig. 3.2(d). As noted in Chapters 1 and 2, this is an improvement upon the original algorithm with linear interpolation [13] because this LI-MAR has identical data fitting and blending to our methods. This image is not much better than the original image.

As discussed earlier, previous numeric methods discard all metal data, treating these points as outliers. We demonstrate that discarding all metal traces leads to a loss in image quality. Figs. 3.3 and 3.4(a) show the optimal solution image when all metal-containing projection samples are discarded. Fig. 3.3(a) shows the optimal solution of the unconstrained problem described in [20]. No metal is visible here because metal traces are discarded. If we use this as the prior-image with our own sinogram completion, we get the image shown in Fig. 3.3(b). Our metal trace estimation is an improvement over data substitution defined in [20] as discussed in more detail below, but we make this comparison because we wish to compare only the effect of prior-images. In Fig. 3.3(b), the large circular liquid object is distorted in shape, and dark shadows are present in each of the four uniform objects. The same shadows can be identified in Fig. 3.3(a).

The effect of the non-negativity constraint along with the deletion of all



(a) (b)

Figure 3.4: Numeric reconstructions (optimal solutions) with the non-negativity constraint. (a) All metal projection data are discarded and the non-negativity constraint is applied. (b) Our weighting function and the non-negativity constraint, but not the constraint in Eq. 3.7. WW=2500, WL=750 MHU.

metal samples is shown by numeric reconstructions in Fig. 3.4. This experiment demonstrates that there is a loss of image quality when all metal is deleted, and that the non-negativity constraint allows some artifacts to persist, as long as the voxel values do not drop below zero. In Fig. 3.4(a) all metal data is discarded. Consequently, objects 1, 3, and 4 are fused, and there is dark shading in objects 1 and 3. Weighting instead of discarding metal trace data separates the fused objects as shown in Fig. 3.4(b). However, the non-negativity constraint still allows dark shading in object 2 in this image. While it is true that CT image values should not be negative, noise, in addition to metal artifacts, can cause negative values in air. The non-negativity constraint is indiscriminate in that it ignores the sources of negative values, while our constraint anticipates where the difference between the measured data and the forward model should be negative.

3.3.2 Visual evaluation

More test cases are shown in Figs. 3.5 and 3.6 which contain pairs of original and MAR images. In each test case, the MAR image has less severe artifacts than the original image.

- Bag 2 contains a long piece of metal. MAR removes the curved bright artifact

under it and reduces the dark streaks.

- Bag 3 shows a dark streak that lowers amplitude by 200-300 MHU, which could result in objects being split by ATR. The arrow points to a bright artifact, which could result in the water phantom and the sheets below the metal to be merged. The large dark streak, smaller streaks and bright artifact are reduced by MAR.

- Bag 4 has an image with a large amount of metal inside and outside a boom-box. The original image has a dark shading artifact in the uniform object, but the MAR image has reduced shading.

- Bag 5 has some new fine streaks in the uniform object due to the many interpolations needed for the cluster of metal objects at the top and from errors between the scanner and synthetic projections of such small dense objects.

- In Bag 6, the object with fine detail on the right hand side of the suitcase appears to be split by the artifact in the original, but is restored after MAR.

In all cases, we see that large dark artifacts between metal pieces or along the long axes of metal pieces are nearly eliminated along with the bright shadows perpendicular to them. These large artifacts are nearly eliminated while the structure of the contents is preserved, because the prior-image included most of the structures but not the artifacts. The narrower streaks are nearly eliminated simply from the interpolation across the traces of the small metal objects. The metal region may not be well reconstructed because we build our prior-image by de-emphasizing metal projections. Although we copy over the metal voxels, if fine spaces exist between metal voxels, they will not be recovered. MAR algorithms can be expected to degrade metal voxels and their neighborhood when they delete or de-emphasize metal projections.

Although there is an overall improvement with MAR in all images, our algorithm has shortcomings as shown in Fig. 3.6. A pot (indicated by a white arrow) in Bag 7 throws off beam hardening artifact (horizontal streak from the

base of the pot). This is not corrected because the pot was not segmented as metal (its CT density is not high enough, probably because it is thin, made of aluminum, and blurred by the system transfer function). The small water bottle (labeled 1) appears fused to the metal above it in the original image, due to bright, smooth metal artifact. With MAR, this bottle is separated but not well restored. Since it is adjacent to a large piece of metal, when the metal traces are given lower weight, most of the projection samples corresponding to this bottle are also de-emphasized, and in effect, we lose too much data. The larger bottle (labeled 2), however, does not share as much data with the metal object. High amplitude streaks through object 2 are reduced, but it also appears joined to the metal.

Bag 8 contains a laptop and metallic bag parts, which create streaks through the water. After MAR, residual artifact is visible in the water. There is also a loss of resolution along the streaks. The loss of resolution, especially along the streaks, is a phenomenon common to most MAR algorithms [36,38]. During data replacement, interpolation across metal traces blurs edges along the rays unless the edges are perfectly captured in the prior-image. However, some of the edges in this image are due to soft materials and thin layers, and therefore are not preserved in the prior-image.

3.3.3 Quantitative evaluation

Each bag contains some uniform objects that are shown numbered in the original images in Figs 3.2, 3.5 and 3.6. The CT number distributions within these objects were measured. The results are given in Table 3.1. The object numbers in the table correspond to the numbering in the original images. The table reveals that the maximum and minimum CT values are closer to the mean value in the MAR images than in the original images. The SD is smaller in the MAR images than the originals for all but one object (object 1 in Bag 7). The KS2 test-statistic is shown in the table column labeled $KS2$. According to the KS2 test, the CT number distributions are different at the 0.05 significance level (p-values not shown). When we consider KS2 along with the SD scores, the KS2 scores tell us that the lower SD of MAR images is not caused simply by a few CT

values in the tails of the two distributions. A few uniform objects were scanned separately. Their means were measured and are given in Table 3.1. In the bags, means shift due to metal artifacts and due to clutter. MAR brings the means closer to the ideal values.

Table 3.2 shows a comparison of the SDs of uniform objects with our MAR and the IPR methods. The mean SD weighted by object volume from IPR is 128, from IPR+ is 117 and from our method is 87 HU. An important finding not represented by uniformity is the distortion of objects in IPR and IPR+ images as shown in Fig. 3.7. Objects appear to be fused together as a result of their edges being blurred in the prior-image (because all sinogram data containing metal are discarded).

Evaluation with SD alone is insufficient because SD can be lowered just by smoothing. For example, a reconstruction algorithm "$x = 0$" will result in SD being trivially zero. Therefore, we use other measures also. Table 3.3 shows the sinogram-based error [66]. This sinogram-based error is intuitive and measures an overall error. In each bag, the error decreases with MAR. The small difference in Bag 1 is likely due to the fact that there are 27 pieces of metal, so much of the sinogram is excluded. The IPR method must have the best scores by definition, because IPR minimizes the same cost that is measured by this error, i.e., the squared error between the original and synthetic sinograms in the metal-free samples.

Finally, we measure image gradients because the sinogram-based error can suppress errors accumulated through reprojection, and does not fully capture how well resolution is preserved. Table 3.4 shows the gradient-based scores. In each bag, the score decreases with MAR, indicating that the MAR images have less severe metal artifacts [66]. More importantly (since the total gradient has the same limitation as SD), when measured only at the boundary of uniform objects, the score is close to that of the original image, indicating that we do not substantially sacrifice edge sharpness to obtain lower SD. The slightly smaller score may be due to a loss of sharpness, or may reflect that the artifacts, which contribute to the total gradient, are reduced. IPR shows the most loss of edge contrast, which matches the visual assessment.

3.3.4 Further analysis

In this section we discuss points that give some more insight into this MAR method.

The LS solution without regularization is equivalent to FBP reconstruction if adequate sampling is present [67]. Therefore, the artifact amplitude and locations are similar in FBP and LS solutions. The numeric solution is influenced by the regularization term, and the FBP solution is influenced by the reconstruction kernel, therefore the FBP and LS solutions are different in details. When we compute the difference of the numeric solutions, we get an image of artifacts, which we subtract from the original image. The prior-image therefore retains the resolution of the original FBP image, and is less influenced by the regularization term than if the optimal solution were used directly.

We chose exponential weights because attenuation is exponential, the weights are smooth and monotonic, and a one-parameter family of weights was easy to tune. The heuristic nature of our weighting motivates the construction of a constraint. The constraint helps to prevent errors from being pushed elsewhere in the image. Fig. 3.8 shows two images with the weights but without the constraints of Eq. 3.7. By comparing these images with those of Fig. 3.5, we see that it is helpful to constrain the solutions. The regularization controls the sparsity of the image gradient. It reduces secondary artifacts that could arise from abrupt changes in weights. Fig. 3.9 demonstrates the effect of β in Eq. 3.7. The allowance for noise provides some flexibility which is helpful when the constraints are too tight. With the current constraint parameter tuning and noise on this dataset, allowance for noise does not make much difference.

We are agnostic about the sinogram completion technique. However, the technique should be robust enough against small errors in the prior. NMAR [38] interpolates a ratio of the original and synthetic sinograms. When synthetic sinogram samples have small values, this ratio has large errors. Large errors in samples near metal will lead to secondary artifacts. This problem occurs near metallic bag parts of a bag that is not tightly packed, since there is little material to attenuate those rays. Our prior-image thresholding also lowers the amplitude of the

synthetic samples. Fig. 3.10 shows the secondary artifacts. All of the images reconstructed with NMAR have secondary artifacts at least as strong as the image shown. Sinogram pre-processing may help NMAR work with our prior.

We draw a distinction between our method and those that perform sinogram subtraction without a prior-image, exemplified by [68]. The authors find limited improvement in their application and Fig. 3.11 shows this approach is ineffective in our application. Reprojection of metal voxels cannot accurately quantify the metal due to beam hardening itself, as demonstrated by simulations in [69]. Therefore, subtraction of synthetic projections of metal cannot reverse the beam hardening.

Faster optimization is necessary for practical use. The Mosek solver takes about 15-20 minutes to reconstruct the miniature using the low precision setting. NESTA took under 2 minutes for a tolerance of 10^{-4}. Both are general purpose solvers. A special purpose solver will allow us to reconstruct larger optimal images to better preserve small structures. Or it could be used to obtain practical reconstruction times. We discuss optimization strategies in Chapter 4.

3.4 Future work

There are many areas to explore for further reduction of residual artifacts and preservation of small structures. 1) We can improve the constraints or add to them, e.g., box constraints per image voxel. Another example is that instead of thresholds, alternative criteria may give better results for determining which projections to constrain. 2) We should explore the use of alternate objective functions [70–72]. 3) Further research in weighting methods is likely to give better results. We also can explore how to quantify the metal. We used indicator functions on the metal rather than summing the value of the metal. The reconstructed metal itself is degraded and forward projection cannot reliably quantify it. Poor quantification of metal gives rise to unreliable weights. 4) More sophisticated sinogram completion may give better image quality e.g., with variational inpainting. 5) Scatter correction and alternate materials for calibration may improve overall image quality.

Although this method was developed for luggage because of the large amount of metal and unpredictability of contents, we believe this method can be applied to medical imaging. Refinements and parameter tuning may be necessary. The prior-image based MAR methods in medical CT have shown that it is not necessary to preserve soft-tissue detail in the prior-image, but it is necessary to adequately reconstruct bone, air pockets and high contrast interfaces. The soft tissue details are captured in the final reconstruction.

3.5 Conclusions

We have developed a new MAR method and tested it on images of luggage with up to 27 pieces of metal. The results of our method show that metal artifacts were significantly reduced based on visual assessment and quantitative evaluation. Our contributions are in two areas. 1) A new formulation of an optimization problem, including projection weighting and a constraint. Methods applicable for luggage scanning, and many MAR methods in general, discard metal projections, but we do not, so details and contrast are better preserved. We use a constraint that accomodates beam hardening and scatter, and gives better results than the non-negativity constraint of previous literature. 2) The difference of solutions to two different optimization problems removes the effects of mismatched spatial resolution from FBP and optimal solutions, and isolates the artifacts.

This chapter contains material from "Metal Artifact Reduction for CT based Luggage Screening", which was co-authored by Pamela Cosman and Harry Martz and was presented at the *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014*.

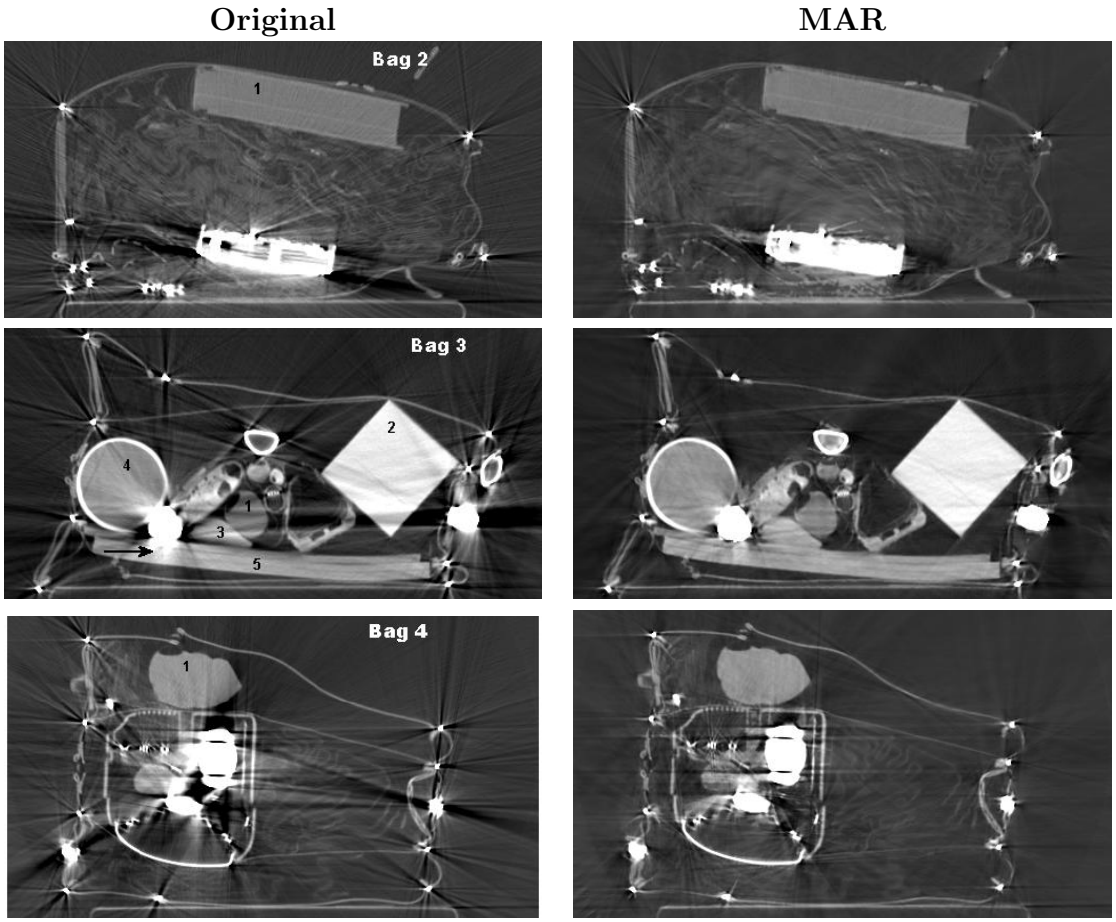


Figure 3.5: Images showing a variety of objects, metals and configurations. Each row shows one test case. Original images are shown on the left, and MAR images are on the right. The numbered objects in the original images are uniform objects that are numerically evaluated in Table 3.1. The black arrow in Bag 3 indicates an example of a bright metal artifact. WL/WW (HU) for Bags 1,2,4,6 = 750/2500, for Bag 3 = 850/2300, and for Bag 5 = 750/2100.

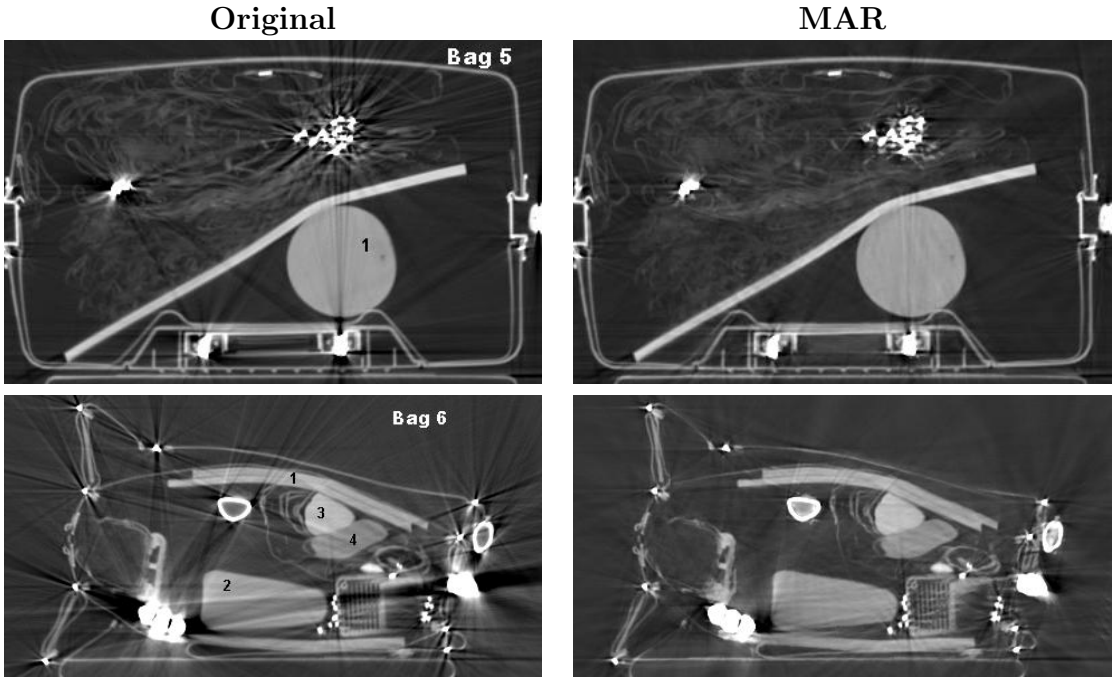


Figure 3.5: Images showing a variety of objects, metals and configurations. Each row shows one test case. Original images are shown on the left, and MAR images are on the right. The numbered objects in the original images are uniform objects that are numerically evaluated in Table 3.1. The black arrow in Bag 3 indicates an example of a bright metal artifact. WL/WW (HU) for Bag 1,2,4,6 = 750/2500, for Bag 3 = 850/2300, and for Bag 5 = 750/2100. (continued)

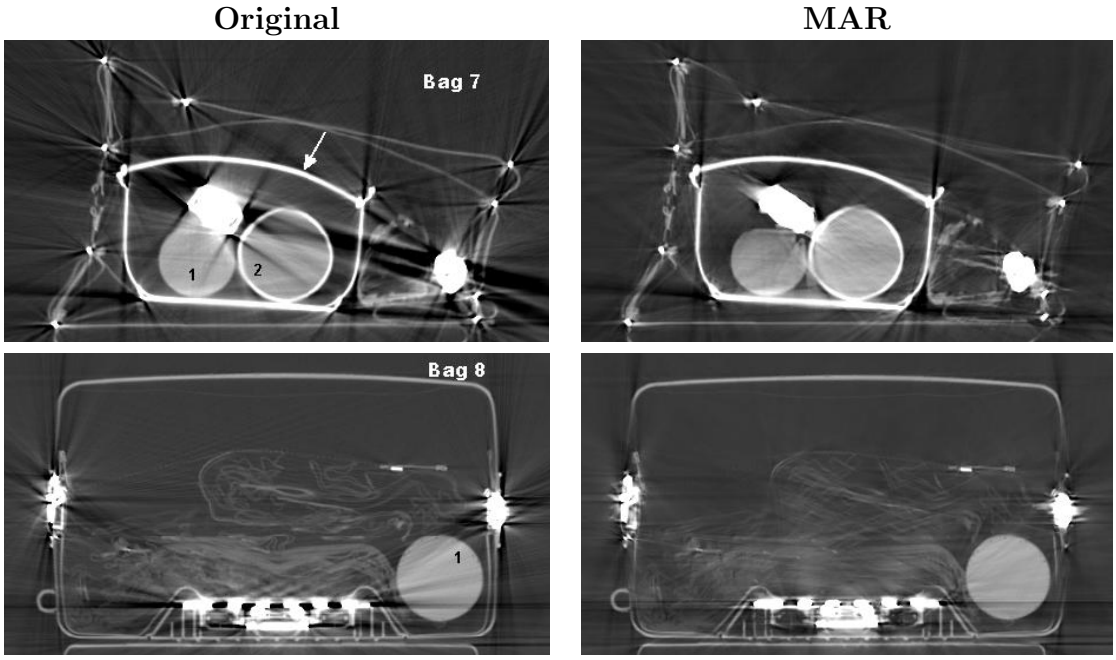


Figure 3.6: Original and MAR images showing some shortcomings of our method. WL/WW (HU) Bag 7: 800/2400, Bag 8: 650/2300.

Table 3.1: The measured CT number distribution in uniform objects. In the original images, these objects (Obj) are numbered 1, 2 etc., and after MAR, they are numbered M1, M2 etc. The SD is lower in the MAR image in all objects but one (Bag 7, object 1). The average SDs weighted by object size are 162 and 87 MHU for original and MAR images respectively. The ideal value marked with * was not available, but is distilled water in a different bottle.

Bag	Obj	Min	Max	Mean	Ideal	Std	KS2
Bag 1	1	155	1296	843		159	0.34
	M1	713	1080	886		48	
	2	380	1140	769		133	0.41
	M2	623	1061	843		65	
	3	582	1457	988		162	0.15
	M3	632	1309	1006		114	
	4	692	1327	1025		79	0.25
	M4	474	1212	979		72	
Bag 2	1	513	850	690		41	0.07
	M1	552	850	687		36	
Bag 3	1	306	1352	695		201	0.46
	M1	569	1116	867		101	
	2	814	2315	1795	1920	164	0.31
	M2	962	2135	1866		74	
	3	718	1868	1276		220	0.20
	M3	959	1460	1260		114	
	4	358	2546	1092	1004	228	0.24
	M4	591	2145	1049		121	
	5	-427	2644	1114		316	0.17
	M5	60	1624	1131		136	
Bag 4	1	456	1155	917		111	0.14
	M1	700	1129	958		65	
Bag 5	1	553	1108	991	1002	60	0.09
	M1	737	1135	998		38	
Bag 6	1	837	1513	1150		106	0.04
	M1	909	1431	1147		94	
	2	288	1600	910	1002*	226	0.23
	M2	371	1357	919		127	
	3	1187	1568	1337		72	0.24
	M3	1186	1529	1356		54	
	4	618	1080	841		90	0.38
	M4	671	1080	907		72	
Bag 7	1	862	1954	1244		143	0.69
	M1	387	1912	992		182	
	2	-205	1437	935		306	0.52
	M2	938	1728	1230		113	
Bag 8	1	269	1229	939	1001	145	0.17
	M1	504	1138	958		94	

Table 3.2: The SD in uniform objects for IPR and IPR+ and our method, respectively denoted Std-IPR, Std-IPR+ and Std-Ours. The objects in each image are numbered as in the original images. The symbols KS2 and KS2+ denote KS2 values that compare our method against IPR and IPR+ respectively.

Bag	**Object**	**Std-IPR**	**Std-IPR+**	**Std-Ours**	**KS2**	**KS2+**
Bag 1	1	96	70	48	0.36	0.09
	2	109	93	65	0.36	0.13
	3	186	128	114	0.34	0.15
	4	181	125	72	0.22	0.14
Bag 2	1	89	82	36	0.30	0.26
Bag 3	1	129	110	101	0.30	0.17
	2	134	111	74	0.20	0.14
	3	134	95	114	0.12	0.30
	4	127	120	121	0.31	0.22
	5	165	149	136	0.19	0.11
Bag 4	1	96	101	65	0.14	0.23
Bag 5	1	58	43	38	0.22	0.08
Bag 6	1	187	143	94	0.40	0.23
	2	143	136	127	0.21	0.10
	3	64	61	54	0.27	0.27
	4	130	124	72	0.26	0.15
Bag 7	1	288	292	182	0.13	0.13
	2	148	177	113	0.14	0.16
Bag 8	1	76	102	94	0.29	0.07
Mean		128	116	87		

Table 3.3: The sinogram-based errors for each bag are smaller after MAR than the original.

Bag	Bag 1	Bag 2	Bag 3	Bag 4	Bag 5	Bag 6	Bag 7	Bag 8
Orig	3.18	4.52	4.26	4.10	4.37	3.74	4.75	5.95
MAR	3.02	3.38	3.16	3.32	4.01	2.83	3.78	3.43
IPR	2.72	2.09	2.75	2.62	3.39	2.84	3.12	3.13
IPR+	2.99	4.46	3.26	3.30	3.81	2.88	3.71	3.53

Table 3.4: Normalized sum of gradient magnitudes. The top three rows contain values computed from the entire image. Lower scores represent better image quality. The bottom three rows contain values computed in the boundary regions. Higher scores represent better image quality.

Bag	Bag 1	Bag 2	Bag 3	Bag 4	Bag 5	Bag 6	Bag 7	Bag 8
MAR	0.83	0.73	0.77	0.81	0.87	0.80	0.80	0.81
IPR	0.93	0.72	0.84	0.86	0.92	0.92	0.93	0.84
IPR+	0.80	0.71	0.73	0.77	0.87	0.77	0.76	0.80
MAR	0.94	0.98	0.92	0.95	0.98	0.95	0.91	0.95
IPR	0.89	0.96	0.86	0.91	0.92	0.91	0.88	0.92
IPR+	0.88	0.96	0.87	0.93	0.94	0.90	0.85	0.92

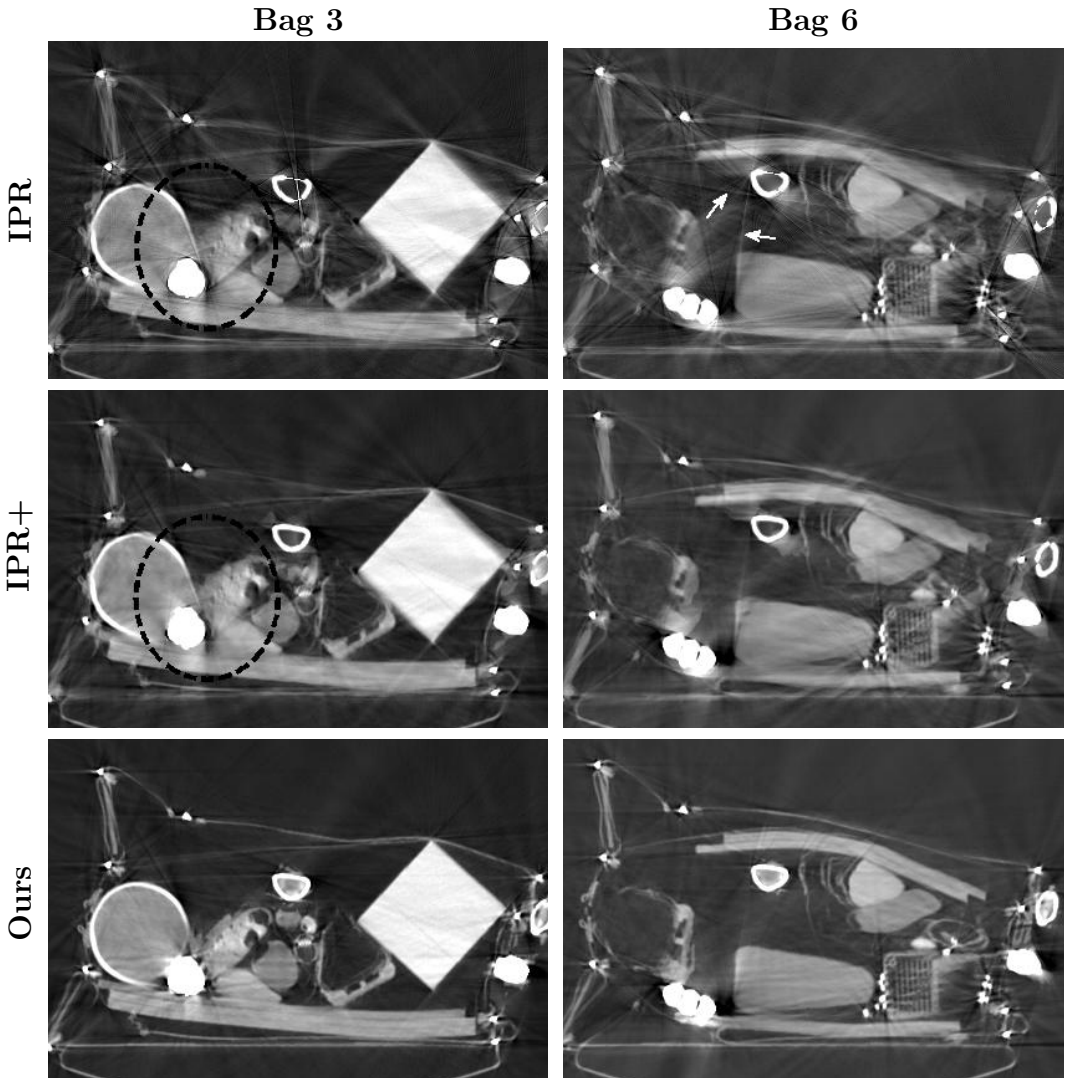


Figure 3.7: Comparison of IPR, IPR+ and our method on Bags 3 and 6. IPR shows a loss of edges e.g., inside the dashed oval, IPR+ shows a small improvement over IPR, and our method shows the best restoration of edges. In the Bag 6 IPR image, arrows point to the streak artifacts from the substitution of reprojected prior-image samples for the original sinogram samples. WL/WW Bag 3: 850/2300 Bag 6: 750/2500.

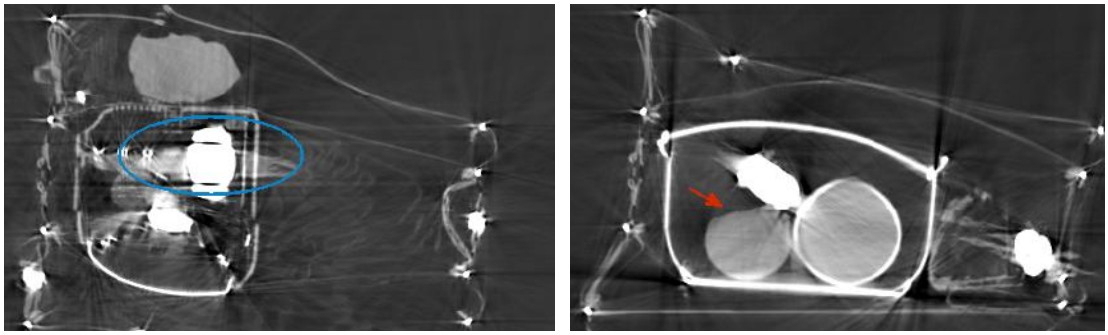


Figure 3.8: More images with the weights but not the constraints of Eq. 3.7. WL/WW =750/2500. The abrupt cutting off of the water is caused by the clipping step in the generation of the prior-image.

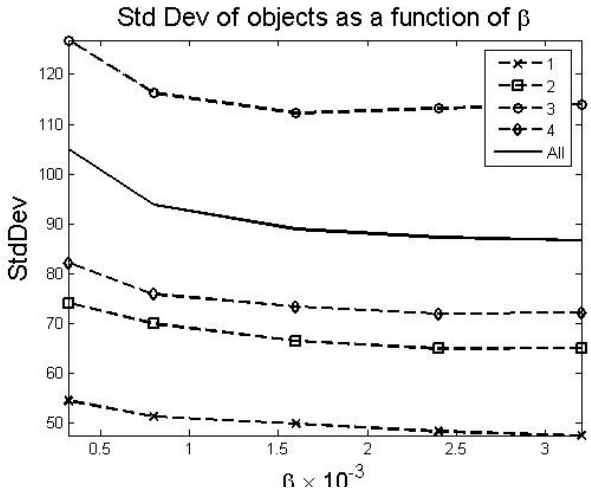


Figure 3.9: The effect of varying β in Eq. 3.7. The first four objects of Bag 1 are shown by lines with markers and the weighted mean of all objects in all bags is shown by the black line.

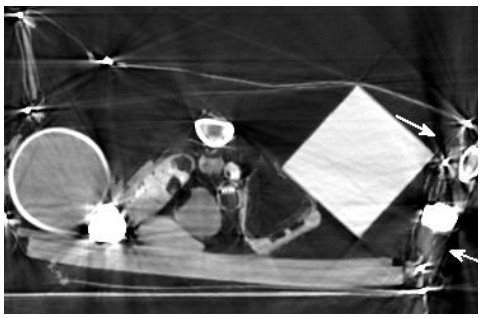


Figure 3.10: Sinogram completion by NMAR. Secondary artifacts are indicated by white arrows. WL/WW = 850/2300.

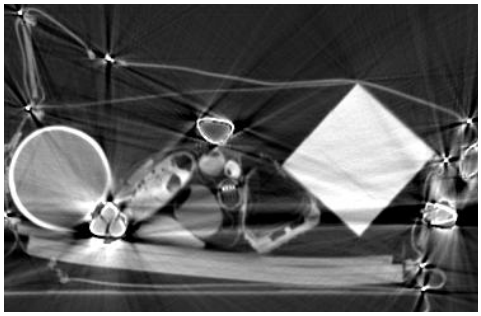


Figure 3.11: Sinogram subtraction without a prior-image. WL/WW = 850/2300.

Chapter 4

Optimization strategies

The Mosek solver is a general purpose solver for linear, quadratic, and SOCP problems, and is slow. Mosek takes about 15-20 minutes to find the solution of our miniature convex problem Eq. (3.7), where the size of A is 46,080×16,384. In this chapter, we present some optimization strategies for faster solutions. The first strategy is based on the alternating direction method of multipliers (ADMM). The second strategy is based on algebraic reconstruction technique (ART).

4.1 Alternating direction method of multipliers solution

The principles of ADMM are explained in [73]. When an objective function consists of separable terms, the terms can be minimized alternately, using the augmented Lagrangian. Consider the objective function below.

$$\begin{aligned} \min_{u_1,u_2} \quad & f(u_1) + g(u_2) \\ \text{s.t.} \quad & Au_1 + Bu_2 = C \end{aligned} \tag{4.1}$$

The augmented Lagragian is defined as

$$L_\rho(u_1, u_2, y) = f(u_1) + g(u_2) + y^T(Au_1 + Bu_2 - C) + \frac{\rho}{2}||Au_1 + Bu_2 - C||_2^2 \tag{4.2}$$

The ADMM solution is

$$\begin{aligned} u_1^{k+1} &:= \operatorname{argmin} L_\rho(u_1, u_2^k, y^k) \\ u_2^{k+1} &:= \operatorname{argmin} L_\rho(u_1^{k+1}, u_2, y^k) \\ y^{k+1} &:= y^k + \rho(Au_1^{k+1} + Bu_2^{k+1} - C) \end{aligned} \tag{4.3}$$

Our objective function consists of the WLS term and the total variation term. The WLS term is minimized by gradient descent, while the TV term is minimized by shrinkage. The minimization is done in alternating steps. Due to our constraint that depends on a weighted sum of the variable x, we add a variable u_1 for using a substitution of Ax. The shrinkage problem for the L_1 norm of the derivative is made easier when we use another subsitution variable $u_2 = Dx$:

$$\begin{aligned} \min_{x,u_1,u_2} \quad & \tfrac{1}{2}(u_1 - b)^T W(u_1 - b) + \beta||u_2||_1 \\ s.t. \quad & u_1 = Ax \\ & u_2 = Dx \end{aligned} \tag{4.4}$$

The augmented Lagrangian for the ADMM problem is written as

$$\begin{aligned} L_\rho(x, u_1, u_2, y_1, y_2) = \; & (u_1 - b)^T W(u_1 - b) + y_1^T(u_1 - Ax) + ||u_2||_1 + y_2^T(u_2 - Dx) \\ & + \frac{\rho_1}{2}||u_1 - Ax||^2 + \frac{\rho_2}{2}||u_2 - Dx||^2 \end{aligned} \tag{4.5}$$

The x-subproblem is

$$x = \operatorname*{argmin}_x \frac{\rho_1}{2}||u_1 - Ax||^2 + \frac{\rho_2}{2}||u_2 - Dx||^2 - y_1^T(u_1 - Ax) - y_2^2(u_2 - Dx) \tag{4.6}$$

We choose gradient descent because we do not want to invert the matrix A. A gradient descent update is defined as

$$x^{k+1} = x^k - \tau\nabla_x(L) \tag{4.7}$$

The derivative of the Lagrangian with respect to x is

$$\nabla_x(L) = (\rho_1 A^T A + \rho_2 D^T D)x - A^T(\rho_1 u_1 - y_1) - D^T(\rho_2 u_2 - y_2) \tag{4.8}$$

Our gradient descent step is therefore defined as

$$x^{k+1} = x^k - \tau((\rho_1 A^T A + \rho_2 D^T D)x^k - A^T(\rho_1 u_1 - y_1) - D^T(\rho_2 u_2 - y_2)) \tag{4.9}$$

The u_1 subproblem is

$$\begin{array}{ll} u_1(i) = \underset{u_1}{\text{argmin}} & \frac{1}{2}(u_1 - b)^T W(u_1 - b) - y_1^T(u_1 - Ax) + \frac{\rho_1}{2}||u_1 - Ax||^2 \\ \text{s.t.} & I_p(u_1 - b) \succeq 0 \end{array} \tag{4.10}$$

The u_1 sub-problem can be minimized by taking the derivative because the matrix inversion is easy. We do not have to use gradient descent. Therefore,

$$u_1 = (W + \rho_1)^{-1}(y_1 + \rho_1 Ax + Wb) \tag{4.11}$$

The constraint is handled by projecting u_1 onto the set $u_1|I_p(u_1 - b) \succeq 0$. More specifically, let $\Omega = \{i|[I_p]_{i,i} = 1\}$ be the index set. Then, for any $i \in I_p$, we let

$$[u_1]_i = \begin{cases} [u_1]_i & [u_1]_i \geq b_i \\ b_i & otherwise \end{cases} \tag{4.12}$$

The u_2 subproblem is

$$u_2 = \underset{u_2}{\text{argmin}}\, \beta||u_2||_1 - y_2^T(u_2 - Dx) + \frac{\rho_2}{2}||u_2 - Dx||^2 \tag{4.13}$$

This can be solved by the shrinkage formula [74]:

$$u_2 = \max\left\{|Dx + \frac{y_2}{\rho_2}| - \frac{\beta}{\rho_2}, 0\right\} \cdot \text{sign}(Dx + \frac{y_2}{\rho_2}) \tag{4.14}$$

The updates to y_1, y_2 are the usual updates for ADMM problems:

$$\begin{array}{ll} y_1^{k+1} & = y_1^k - \rho_1(u_1 - Ax) \\ y_2^{k+1} & = y_2^k - \rho_2(u_2 - Dx) \end{array} \tag{4.15}$$

The parameter values, given in Table 4.1, were selected by trial and error.

Table 4.1: Parameter values for the ADMM-based solver.

Parameter	**Value**
ρ_1	0.02
ρ_2	0.04
τ	5×10^{-4}
β	20

The image X^C_{mini} was reconstructed using this method, and it took about four minutes to run about 5000 iterations in C code, on an Intel i7 computer with

3GB memory. A reconstructed image is shown in Fig. 4.1. The SDs of uniform objects in this image are given in Table 4.2. The SD values are higher than the Mosek-based images. However, this may be due to our parameter selection. Better parameter selection could lead to better image quality, convergence in fewer iterations, or both. The area around the metal itself has been reconstructed at least as well as the Mosek-based solution.

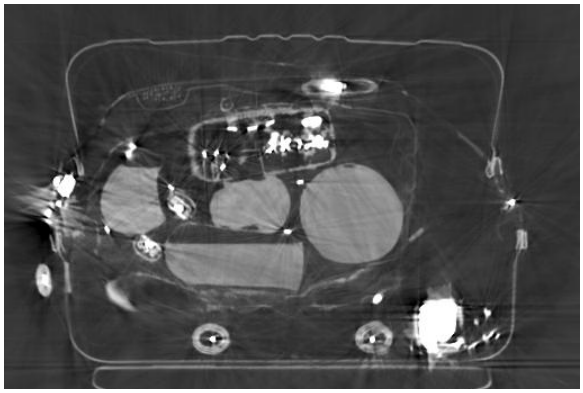


Figure 4.1: MAR image using an ADMM-based prior-image.

Table 4.2: Table showing the measurement statistics for the image reconstructed with the ADMM solution for X^C_{mini}.

Object	**Min.**	**Max.**	**Mean**	**SD**	**KS2**
M1	686	1122	884	61	0.29
M2	593	1094	836	85	0.32
M3	455	1360	1011	121	0.15
M4	520	1262	982	80	0.21

4.2 Algebraic reconstruction technique (ART)

The method of Kaczmarz, which is a projection onto convex sets (POCS) method, is known as ART in CT reconstruction. It is defined by:

$$x^{k+1} = x^k + \beta^k \frac{(A_i^T x - b_i)}{||A_i||_2^2} A_i^T \ , \tag{4.16}$$

where A_i is the row of the matrix for sample i, and k represents the current update. The samples are chosen according to $i = k \mod M$ where M is the number of

samples, i.e., rows in A. In this equation β^k is a relaxation parameter, which can be decreased according to some schedule.

ART was one of the earliest reconstruction methods used in CT. As a statistical method, it could incorporate various characteristics of the scanning system by modeling the A matrix. It was replaced by FBP from preprocessed projection data. FBP is much faster but expects ideal data.

Using preprocessed data in ART, like that used in FBP, reduces the burden on modeling A, and improves the convergence. In recent years, ART has been used in reconstruction from limited data [49]. In this adaptation, called adaptive-steepest-descent-POCS (ASD-POCS), the objective function is the total variation norm. This objective function must be minimized such that the constraints $||Ax - b||_2 < \delta$ are satisfied. Non-negativity is imposed after every iteration of ART in [49]. The minimization of the objective function is achieved by steepest descent calculation of the total variation norm. As mentioned in Section 1.2.3, this reconstruction algorithm has been used to reconstruct from sinograms after discarding metal trace data.

The POCS approach does not use weighting. If the A, b were modified in Eq. 4.16, it would have no effect on the solution x. We have modified ART by adding a step-size w_i. The step size decreases with the amount of metal in the projection. We set w_i to the diagonal elements of $W_{X_{Orig}}$.

$$\begin{aligned} x^{k+1} &= x^k + w_i \beta^k \frac{A_i^T x - b_i}{||A_i||_2^2} A_i^T, \quad [I_p]_{i,i}(A_i^T x - b_i) \geq 0 \\ &= 0, \qquad\qquad\qquad\qquad \text{otherwise} \end{aligned} \tag{4.17}$$

A MAR image with X^C_{mini} reconstructed using Eq. 4.17 is shown in Fig. 4.2. It was reconstructed in approximately four seconds. The SD measurements are in Table 4.3.

4.2.1 Fast projector-backprojector pair

The optimization problem described thus far builds miniature images. For larger problems, the system matrix becomes impractical to store in memory. We replace the matrix-vector multiplications with operators so that we do not have

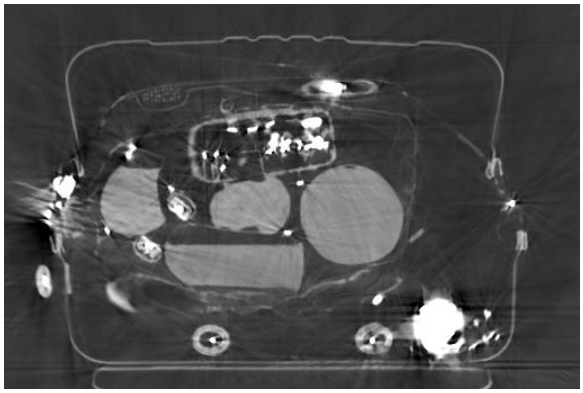


Figure 4.2: MAR image using an ART-based reconstruction for X^C_{mini}.

Table 4.3: Table showing the measurement statistics for the MAR image using the ART-based solution for X^C_{mini}

Object	**Min.**	**Max.**	**Mean**	**Std.**	**KS2**
M1	710	1064	902	50	0.36
M2	619	1055	838	71	0.38
M3	661	1320	1015	120	0.16
M4	493	1314	996	77	0.17

to store a large matrix. However, an operator-based algorithm is slower than multiplying stored values. The operator for Ax is forward projection and for A^Tb is backprojection.

We have developed a fast projector-backprojector pair. Considering the ART equations, the innermost operation begins with forward projection for some ray i. In our method, we locate the image voxels that the ray intersects. We store each voxel position, and weights that represent how much each voxel contributes to that ray. We sample the ray at discrete points, and obtain real number spatial coordinates in image voxel units. At a given sampling point along the ray, we find the four voxels nearest the co-ordinates, and calculate weights equal to bilinear interpolation coefficients. At the next intersection point, we get another four voxels and their weights.

Our sample spacing is equal to the voxel spacing, so consecutive sampling points intersect one, two or four common voxels. We would like to store the voxels without repetition, so that fewer multiplications are required during backprojec-

tion, and less storage is needed for each ray. We solve this problem by using a linked list with a voxel storage order that depends on the angle of the projection ray. An illustration of the storage order is in Fig. 4.3 for one angle. The diagram shows image voxels as gray nodes on a grid. Two parallel projection rays, that intersect different voxels, are shown in the diagram. The colored arrows show the order in which the image voxels are stored for a given ray.

Ray 1 is sampled at the equi-spaced points marked P1 - P3. The four voxels around P1 are stored in the order upper left, lower left, upper right, and lower right. These voxels are labeled 1-4 in the diagram. P1 is followed by P2 such that the set of voxels for P2 is one column away from the set for P1. Two of the voxels are common, and we continue the order of the voxels, by adding only the upper right (5) and the lower right (6) to the stored set. The next sample, P3, is such that its voxel set is located in the next column and the next row. Again, the same storage order can be followed, adding the lower left (7), followed by upper right (8), and then lower right (9). This order is applicable when a new sample point has the same voxel set, a set one column away, or one column and one row away from the previous sample point.

Ray 2 shows a different sequence. Again, the samples are called P1-P3, and the order for the voxel set for P1 is 1-4. P2 was represented as before: upper left, lower left, upper right, lower right because its voxel set was to the right of the set of P1. When P3 is sampled, the storage order is changed. The voxel set of P3 is one row below that of P2. The lower left voxel for P2 no longer points to the upper right (green arrow). Instead it now points to the lower left voxel for P3, which in turn will now point to the upper right for P2, as shown by the long red arrow. These are now voxels 5 and 6. The link between upper right and lower right neighbors for P2 was never disrupted when P3 was added. The lower right neighbor for P2 is the upper right neighbor for P3, which points to the lower right for P3, as usual, i.e, points 7 and 8. Therefore, only one link was replaced.

The zig-zag order we have described is different for projection rays depending on their angle. We define separate orders for rays in the intervals [0 - $\frac{\pi}{4}$), [$\frac{\pi}{4} - \frac{\pi}{2}$), [$\frac{\pi}{2} - \frac{3\pi}{4}$), [$\frac{\pi}{4} - \pi$), that are mirror images of the illustrated order.

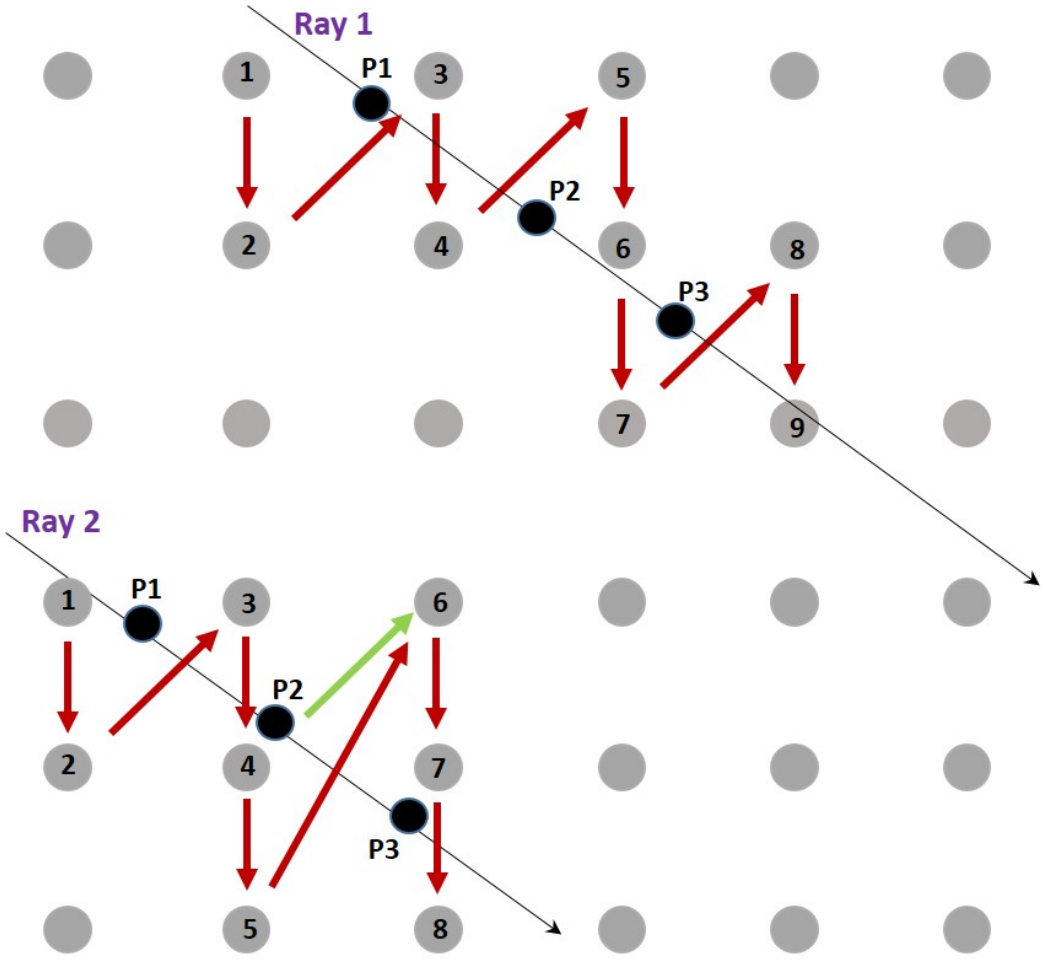


Figure 4.3: Diagram of fast projector/backprojector pair.

We reconstruct a quarter-sized image, i.e., 256×256 voxels. This represents a reduction by a factor of two in each dimension, rather than by four, as with all our previous experiments. The sinogram is also downsampled by two instead of four. We use the reconstructed image directly as the prior-image, instead of obtaining the prior-image by the difference method, as with all our previous work. We do not use the difference method, because when the quarter-size image is upsampled by two, it still retains edge information adequately, as shown in Fig. 4.4. The quarter-sized problem also has less model mismatch from sampling errors. This image takes about 10 minutes to reconstruct. The SD measurements are in Table 4.4.

The SDs of the image reconstructed with a quarter-sized prior-image are

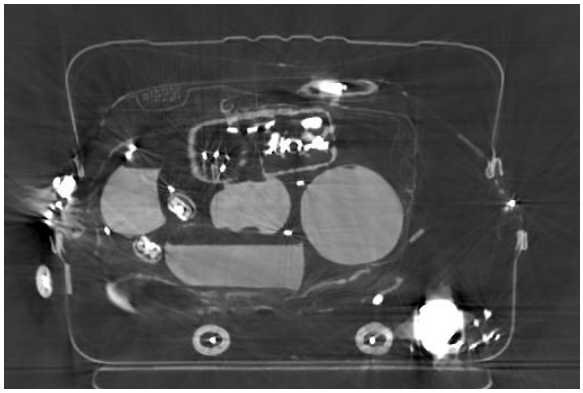


Figure 4.4: MAR image using a prior-image that is a quarter-size of the original image, reconstructed with operator-driven ART.

Table 4.4: Table showing the measurement statistics of the MAR image, using a quarter-sized prior-image, reconstructed by the operator-based ART method.

Object	**Min.**	**Max.**	**Mean**	**Std.**	**KS2**
M1	711	1067	894	45	0.36
M2	404	1074	841	70	0.41
M3	746	1265	1017	99	0.20
M4	716	1165	989	64	0.22

lower than other images. This may be due to the fact that we used cubic interpolation for upsampling the artifact image X_A, which introduces high-frequency errors into the prior-image. Using bilinear interpolation may have been more appropriate to upsample X_A. The ADMM-solution has the best reconstruction of the region around the large piece of metal. This region is as good as that of the Mosek-based image. The ART-based methods do not do as well in this region.

Chapter 5

Evaluation methods for CT segmentation algorithms

5.1 Introduction

The main difficulties for segmentation in luggage screening are the variety and heterogeneity of non-threat and threat objects found in bags, as well as image artifacts. These difficulties cause segmentation algorithms to split an object into multiple pieces, or to merge different objects into a single one.

Quantitative evaluation of segmentation algorithms is a challenging task in luggage screening because multiple splits and merges are possible. In addition to an accuracy score, we would like to gain a deeper understanding of the algorithms' behavior. First, we would like to know if an algorithm systematically oversegments or undersegments images or if the error is random. A knowledge of systematic errors allows us to tune the parameters of a segmentation algorithm, or supplement the segmentation algorithm with additional steps such as region merging [52]. Second, the ability of a segmentation algorithm to capture object features must be evaluated, because evaluation of features is critical in ATR. Third, since it is often more important to correctly segment some objects than others, a method to assign priorities to segments is desirable when evaluating the algorithm. Priorities may be based on image intensity, homogeneity, particular texture or any other

image features that define objects of interest. Fourth, a segmentation algorithm may have varying accuracy across the feature range, and this knowledge can be used to establish confidence in a given segment. There can be no restriction on the number or nature of objects. All these considerations are important in luggage scanning but are not adequately addressed by existing evaluation literature.

Various goodness measures have been proposed to evaluate a segmentation without a ground truth (GT) [75–77]. The goodness measures are based on entropy, intra-region similarity and inter-region discrepancy, surface smoothness and other properties of regions. However, objects found in luggage are inherently heterogeneous, i.e., made up of different materials that have different textures and attenuation properties. Their sizes and shapes are varied and unpredictable. Therefore, goodness measures are not well-suited to our problem.

There are many methods that evaluate segmentation against GT by computing a distance between the sets of edge pixels [75, 78, 79] or surface voxels [80]. However, edge or surface distances do not measure feature retrieval. Mass or volume may be well retrieved, but have large edge distances due to artifacts or other segmentation errors. Therefore, using discrepancy between sets of edges does not appear to be a good solution for luggage screening.

An error measure was defined to measure the discrepancy among manual segmentations performed by multiple humans. This measure was designed to be unaffected by refinements [81]. For the purpose of human perception from photographs, refinements were considered alternative and equally good GTs. For example, a container and its lid may be segmented together or separated by a contour. In the luggage application, we have object splitting and merging instead of refinements. These splits or merges cannot be considered alternate ground truths, but rather errors that must be measured. Therefore, we tested another error measure created with the objective of quantifying the splitting and merging, called the object consistency error (OCE) [82]. OCE is sensitive to refinement. We found that OCE does not perform well with splits and merges that are not simple refinements. This issue is illustrated in Section 5.4 using synthetic examples.

Another method breaks down the evaluation problem into the measurement

of correct detection, oversegmentation, undersegmentation, missing objects, and spurious detection [83]. However, the method depends on the existence of planar surfaces in the image. The measures discussed in [84] interpret the different labeled regions as clusters, and measure distance between clusterings. The wide range of the number of labels from the different machine segmentation (MS) algorithms makes it unsuitable to apply pair-wise clustering interpretations to the labels (such as the Rand index [85]) because some of the labels have small cardinality.

The segmentations may be viewed as different partitionings. A metric was defined as the minimum number of elements that must be moved from one partitioning in order to get to the other partitioning [86]. In these methods, no calculation of systematic errors, no feature-based evaluation and no assignment of priorities among partitions was described, which is needed for our application and may be important in others. Other single-valued measures include [87–89].

The evaluation methods cited above are based on volume or surface overlap. However, there are other features of objects that are more relevant for ATR than volumes or surfaces. A measure called ultimate measurement accuracy (UMA) computes a distance between the measurements of a feature made in the GT and MS images [90]. This measure works on single foreground objects. A multidimensional evaluation is in [91] but systematic errors are not quantified. The treatment of segmented images as probability mass functions was suggested for a 2-class problem [92]. The divergence measure is similar to the Kullback-Leibler (KL) divergence. The idea was improved upon, to measure features of collections of similar objects by creating histograms of feature values for populations of similar objects, and comparing them using standard histogram comparison measures [93]. Similarly, image-distance measures were suggested in [94] based on feature distribution similarity.

We propose two new methods of evaluation to meet the application needs described above, and address many limitations of existing methods. In the first method, we calculate a weighted mutual information (WMI) of features from their joint distribution. In the second method, which we call feature descriptor recovery (FDR), we measure systematic and overall errors in feature recovery, and extract

additional information about behavior over feature ranges. The two methods provide different evaluation perspectives. They are flexible in that they can operate on features, they impose no restrictions on the type or number of segmented objects, and can prioritize segments by feature values or user preference.

In luggage screening, air, which occupies a large portion of the images, is not segmented. In this project, we treat air differently from other labels so that missing objects are penalized, but spurious objects are not. The spurious objects are bag parts and image quality verification phantoms present in the scans that were labeled by some of the research groups, but were not labeled in the GT. We do not want to penalize (or reward) these spurious objects. Our methods allow us to discard the spurious objects. This is different from the ATR problem, where the spurious objects are analogous to a false alarm. Although we do not specifically discuss false alarms due to the classified nature of threats and ATRs, it is certainly desirable for EDS vendors and testing agencies to assign penalties using appropriate weights within our framework. As we will see later, we cannot treat air as simply another label, because that would allow purely nominal scoring methods to reward missed objects.

We applied our evaluation methods to images from the ALERT dataset. Our evaluation methods were validated (1) by applying them to simple synthetic problems, and (2) by comparing the methods' results on suitcases with an evaluation done by a human observer. In information theory, the F_1 score is an accepted measure of performance for binary classification problems [95]. We have also compared our results against a multi-class generalization of the F_1 score.

5.2 CT images and ground truth

Suitcases were scanned on a volumetric medical CT scanner. The suitcases contain objects such as clothing, shoes, electronics, food, books, toys and various contained liquids. These suitcases do not contain threats or simulants. The CT image dimensions were 512×512 pixels per image slice with about 800 slices in each (3D) image. Segmentation algorithms were developed by five different re-

search groups for this project: Siemens Corporate Research, Marquette University, University of East Anglia, Stratovan Corp, and Tele-Security Systems. Each algorithm was run on five test images and generated a label image (except one image by one research group), so we have a total of 24 MS label images. Further details are in [53]. Descriptions of some of the segmentation algorithms are available for the interested reader in [43, 96, 97].

We developed a computer-assisted method to generate GT label images from volumetric CT images. This method was used by researchers at Northeastern university to generate ground truths from the suitcase images. Objects found in luggage are so varied in size and shape, and heterogenous in material composition, that human interaction is required to define the GT. However, segmentation performed solely by a human is limited in accuracy by the observer's ability to manually contour objects. The objects are not only three-dimensional, but sometimes hollow or thin and with large surface areas. Further, the objects are blurred by the transfer function of the CT scanner. For all these resasons, it is impractical to segment these objects by an exclusively manual method. In our method, manual contouring is complemented with manually-seeded region-growing, allowing the observer to segment complicated shapes. Manual segmentation by multiple observers has been addressed by [98, 99]. However, our challenge comes not from subjective perception, but rather from the difficulty of the manual task, which necessitates some automation. Another 3D GT generation method [100] uses mesh models. Our approach is simpler and does not require modeling.

A unique label is assigned to each object that is individually packed into the suitcase. For example, a liquid is assigned the same label as its container. While the validity of this rule may be argued, it overcomes the issue of subjective perception. There is no soft (probabilistic) label assignment for image voxels. A label value of zero indicates air, which is background. Objects with an average CT value of less than -500 Hounsfield units (HU), such as clothes, were assigned the label of air. In the HU scale, air is -1000 and water is zero.

Each bag image file contains hundreds of image slices. However, the observer does not have to contour objects in every slice. Interpolation is performed

between contours across slices, allowing the GT segmentation to be completed in roughly two hours per bag. The interpolated contours are filled in and labeled as one object. Objects in a bag are segmented one at a time, given unique labels and accumulated into a GT label image. Fig. 5.1 shows a flowchart of the computer-assisted GT extraction process. There are two parallel paths. The right half of the flowchart illustrates the manual contouring, contour interpolation and contour filling processes. The output of this path is an image of a filled three-dimensional contour, labeled A in the flowchart. The left half illustrates the seeded region-growing path. The user selects seeds and region growing parameters for the current object. The output of this path is a region-grown mask in image B. The common voxels in images A and B are assigned a user-selected label value Λ. The label image for this object is accumulated with previously segmented labels in image C. We use the maximum operation instead of a binary "or" operation to set a rule for overlapping labels. This rule is useful when there are touching objects. If the first object is labeled as Λ_1, as the user segments the second object, the two labels may overlap. The user resolves the problem by selecting $\Lambda_2 < \Lambda_1$ for the second object's label if he decides that the common voxels should belong to the first label or $\Lambda_2 > \Lambda_1$ if they should belong to the second one.

The flowchart in Fig. 5.1 was implemented in MeVisLab [101], a graphical programming language that provides image processing and visualization modules that can be connected together. In our program, DICOM images are read in and multi-planar reformatted (MPR) slices are generated and displayed. The user selects an MPR axis and contours an object in any slices along the chosen axis. We used a drawing tool that incorporates active contours. The active component helps the contour to be attracted to gradients or curvature as determined by user-defined penalties. A contour can be copied and pasted to other slices. The contours are linearly interpolated to all slices between the first and last contour. The observer subjectively decides whether the contour interpolation provides acceptable results. If the results are not acceptable, more contours can be added and interpolation repeated. The interpolated contours are filled in with a user-selected label value. Binary dilation is performed to include edge voxels.

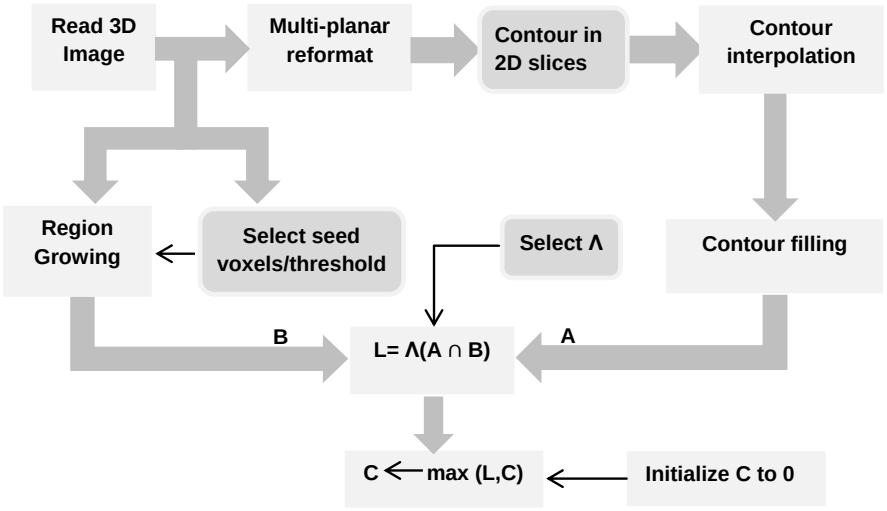


Figure 5.1: Flowchart showing the operations performed to determine GT labels from the volumetric CT image. Manual operations are shown in darker boxes. A and B are segments generated by the two different paths, C is the accumulated set of labels, and Λ is a numeric value assigned to a label.

In the second parallel path, the user selects seed voxels from the object, and sets upper and lower thresholds for region growing. The user can overlay the region grown mask on the CT image to decide whether the mask is acceptable, and modify seeds or thresholds if deemed necessary to repeat the region growing process.

5.3 Segmentation evaluation methods

We first describe the weighted mutual information score for volume-based evaluation, and then describe our extension to mass-based evaluation. Next, we show the weighting functions that allow us to prioritize objects. Then we describe the FDR method which gives systematic errors and total error. Feature descriptors may also be weighted to prioritize objects. Finally, we describe the multi-class extension of F_1 scores, which we use for comparisons.

Let G and S denote the number of labels in the GT and the MS images, respectively, not including the air segment (label 0). Let $\mathcal{X}_G(i)$ be the set of voxels in GT segment i, and $\mathcal{X}_S(i)$ be the set of voxels in MS label i. We use the terms

segment and label interchangeably.

5.3.1 Weighted mutual information (WMI)

Mutual information (MI) can be used when the label images are expressed as joint and marginal probability densities [102]. We generate a confusion matrix from the GT and MS images. We first compute the MI and entropies without air so that MI, which is ordinal, does not reward the air label. Then we include the type II errors with a multiplicative factor. We neglect type I errors for the reasons explained in Section 5.1. Let $N_{G,S}(i,j)$ denote the number of voxels that belong to GT label i and MS label j.

Let $v_{G,S}(i,j)$ denote the joint probability mass function (pmf) based on volume:

$$v_{G,S}(i,j) = \frac{N_{G,S}(i,j)}{\sum_{k=1}^{G}\sum_{l=1}^{S} N_{G,S}(k,l)}, 1 \le i \le G, 1 \le j \le S \ . \tag{5.1}$$

We define the marginal probability mass functions for the GT and MS labels from the joint pmf.

$$v_G(i) = \sum_{j=1}^{S} v_{G,S}(i,j), 1 \le i \le G \tag{5.2}$$

and

$$v_S(j) = \sum_{i=1}^{G} v_{G,S}(i,j), 1 \le j \le S \tag{5.3}$$

A normalized MI score is generated in the following manner, as first described in [103]:

$$H = \frac{1}{Z}\sum_{i=1}^{G}\sum_{j=1}^{S} v_{G,S}(i,j)\log\frac{v_{G,S}(i,j)}{v_G(i)\, v_S(j)} \ , \tag{5.4}$$

where the normalization factor Z is the square root of the product of entropies, or the GT entropy if the MS entropy is zero:

$$Z = \begin{cases} \sqrt{\sum_{i=1}^{G} v_G(i) \log\left(\frac{1}{v_G(i)}\right) \sum_{j=1}^{S} v_S(j) \log\left(\frac{1}{v_S(j)}\right)}, & \text{if } S > 1 \\ \sum_{i=1}^{G} v_G(i) \log\left(\frac{1}{v_G(i)}\right), & \text{otherwise.} \end{cases} \tag{5.5}$$

We now incorporate type II errors. Referring to Fig. 5.2, we take the ratio of the total voxels in the inner matrix (dark shaded) to the total voxels in the outer matrix (all shaded).

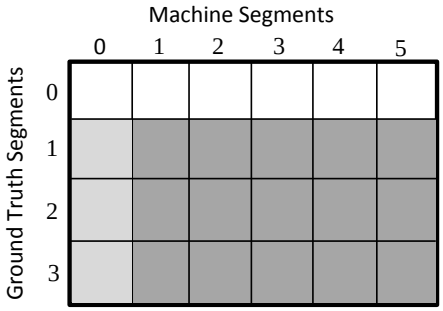


Figure 5.2: Confusion matrix showing inner matrix used in the calculation of entropies, and showing the outer matrix used in r (Eq. 5.6).

$$r = \frac{\sum_{i=1}^{G} \sum_{j=1}^{S} N_{G,S}(i,j)}{\sum_{k=1}^{G} \sum_{l=0}^{S} N_{G,S}(k,l)} \tag{5.6}$$

This ratio is analogous to recall in a binary classification problem, if all objects were considered to belong to one class and air was considered the second class. Recall is also called sensitivity or true positive rate. The unshaded row contains type I error. We multiply this ratio, called r, with H. To make this factor more general, an additional weight can be used so that missed data receive larger or smaller penalty. Our WMI score is given as

$$I = r \times H. \tag{5.7}$$

We chose mass as a basis for evaluation because it gives us information that is different from volume in most luggage articles, but just as relevant to the

application. For example, in shoes, the upper is responsible for most of the volume, while the sole is responsible for most of the mass. The upper is less relevant in the ATR field because of its low density. We now use WMI to measure mass-based score, and then prioritize objects by weighting the confusion matrix. For the mass score, a confusion matrix cell contains not the number of voxels common to a pair of GT and MS labels, but rather, the common mass, which is calculated by summing values from the CT image. Let the CT image be denoted C. Then the mass in cell (i, j) is given by

$$M_{G,S}(i,j) = \sum_{x \in \mathcal{X}_G(i) \cap \mathcal{X}_S(j)} C(x). \tag{5.8}$$

The joint and marginal pmfs for the mass-feature are computed from the confusion matrix in a manner similar to that shown for volume. The mass WMI score can be considered a weighted volume score, with weights equal to the CT number being assigned to each voxel.

Note that these calculated values of volume and mass should be multiplied by voxel size and CT scaling factors to obtain true volume and mass, but these multiplicative factors are constants and can be neglected. Aside from the constant scaling factors, the mass is not the true physical mass of the object, but an approximation. The CT image intensity is proportional to the material linear attenuation coefficient, which itself is proportional to the physical density of the material if we neglect the atomic number of the material and the energy-dependence of the attenuation coefficient.

Next we describe using a weighted confusion matrix (Section 5.3.2) to weight objects and errors. Specifically, we demonstrate weighted mass WMI, and uniformity (a regional feature).

5.3.2 Weighted confusion matrix (WCM)

We assign priorities to segments by weighting the cells of the confusion matrix before computing WMI scores. We define weights that assign greater importance to homogenous (also called uniform) objects. Uniformity is not a feature of interest in ATR, but as we will show later, it demonstrates interesting behavior

of the segmentation algorithms. The mass confusion matrix rows were weighted by a measure of uniformity. This measure of uniformity can be considered a texture feature, weighted by mass to prioritize heavier objects. Alternately, it can be considered a mass feature, weighted to prioritize homogenous objects.

$$w^{\sigma}_{G,S}(i,j) = \frac{1}{\sigma_G(i)}, \tag{5.9}$$

where $\sigma_G(i)$ is the standard deviation of the CT numbers in GT label i and is given by

$$\sigma_G(i) = \sqrt{\frac{1}{N_G(i)} \sum_{x \in \mathcal{X}_G(i)} \left(C(x) - \overline{C_G}(i)\right)^2}, \tag{5.10}$$

where $N_G(i)$ denotes the number of voxels in GT segment i. In the above equation the mean CT number is given by

$$\overline{C_G}(i) = \frac{M_G(i)}{N_G(i)}, \tag{5.11}$$

where $M_G(i)$ is the mass within the GT label i given by

$$M_G(i) = \sum_{x \in \mathcal{X}_G(i)} C(x). \tag{5.12}$$

A natural extension of this idea is cell-wise weighting. We also defined cell-wise weights with the goal of assigning non-uniform costs to different classification errors as shown below. The weights of the cells are lower if they are from dissimilar objects:

$$w^{cell}_{G,S}(i,j) = \frac{\min(\overline{C_G}(i), \overline{C_S}(j))}{\max(\overline{C_G}(i), \overline{C_S}(j))} \tag{5.13}$$

The cellwise weights were applied to the volume confusion matrix $v_{G,S}(i,j)$.

5.3.3 Feature descriptor recovery (FDR)

The FDR method measures how well the features of each object are recovered. Feature descriptors have more flexibility than the WMI framework because label-wise features can be used. For example, one can use label-averages or inter-label separation divided by intra-label uniformity. Weighting can be incorporated, similar to the WMI method.

The numbering of labels in the GT and MS images is arbitrary. We establish the optimal one-to-one correspondence between the GT and MS labels. We used the Hungarian method [104] to maximize the total volume overlap between GT labels and MS labels. Instead of the volume intersection, another cost function could have been minimized, such as the mass intersection.

A feature descriptor P_G is generated by calculating some feature within each label in the GT. P_G is a vector, such that for each label $1 \leq l \leq G$, $P_G(l)$ is the value of the feature computed with respect to the original CT image, within that label. Another feature descriptor P_S is generated for the MS. The feature descriptors P_G and P_S are generated independently of each other.

Features

Analogous to WMI scores, the features we have used are volume, mass, and uniformity, because of their relevance to ATR. As before, the volume of the label is the total number of voxels within the label and the mass of the label is the summed CT value within the label. Similar to the uniformity for WMI, (Eq. 5.9), we define uniformity as the inverse of standard deviation multiplied by the mass, calculated per label. The uniformity feature is shown below for GT labels. It is also calculated for MS labels.

$$W_G^{\sigma}(l) = \frac{M_G(l)}{\sigma_G(l)}, 1 \leq l \leq G \tag{5.14}$$

This feature is similar but not identical to the uniformity-weighted mass of the WCM which had a row-wise weighting (Eq. 5.9). In the WCM, it was not meaningful to consider the standard deviation of the voxels in a cell because a cell can have a small number of voxels. In the FDR method, there is no weighting corresponding to cell-wise weighting w^{cell} of the WMI.

Feature recovery scatter plots

For each object in the MS image and GT image, we generate features as explained in the previous section. For a feature, we generate a scatter plot of the matched labels of P_S against P_G, and call this a feature recovery scatter (FRS)

plot. In any bag, the number of GT and MS labels may not be the same, so the minimum is plotted. The data from all the bags were combined. As explained in Section 5.4, the slope of the line fitted to the data tells us if there are systematic errors. We used a robust fit to reduce the impact of outliers [105].

Residual errors

In order to compute the residual errors from feature recovery, we applied commonly used error statistics, including Cramer-von-Mises (CVM) [106], Kullback-Leibler divergence (KL) [107] and L_1 error normalized by the sum of GT feature values. The L_1-based score is:

$$R_{L_1} = 0.5\frac{|P_G - P_S|_1}{|P_G|_1} \tag{5.15}$$

Although the FRS plots contain the minimum of the number of labels in the MS and GT, the residual error is computed on the maximum of the number of labels. Where a label does not exist, its feature value is zero. The slope of the fitted line and the residual error together provide the performance result.

Behavior over feature range

In addition to over and undersegmentation, the pairing of segments allows us to investigate how accuracy changes over a feature range, and to identify outliers. We take the sliding average (geometric mean) of the feature ratio of the label pairs. The ratio is that of the larger to the smaller feature value. We plot this mean as a function of the sliding geometric mean of the GT labels. This plot indicates the average feature retrieval error against average feature value.

$$R(i) = \left(\prod_{j=-\frac{n-1}{2}}^{\frac{n-1}{2}} \frac{\max(P_G(i+j), P_S(i+j))}{\min(P_G(i+j), P_S(i+j))} \right)^{1/n}, \frac{n-1}{2} < i \leq \min(G,S) - \frac{n-1}{2} \ . \tag{5.16}$$

Ratios are more meaningful than differences in this computation because of the large dynamic range of the feature. In log-scales, this ratio would be the

absolute value of the difference and the geometric mean would be the arithmetic mean, corresponding to taking a sliding L_1 error. This prevents opposite polarity errors from canceling.

From the FRS plots, we can obtain outliers. For each pair of GT and MS points, we compute the following distance.

$$d(i) = \log\left(\frac{P_S(i)}{P_G(i)}\right), 1 \leq i \leq \min(G, S). \tag{5.17}$$

We fit a normal distribution to the distances and obtain its standard deviation, σ. Points $i : \|d(i)\| > 3\sigma$ are considered outliers.

5.3.4 Multiclass F-score ($\mathbf{F}_1^m$)

In information theory, the F-score is an accepted measure of performance for binary classification problems [95]. We generated a multi-class extension of the F_1 score to help validate and offer some perspective on our results. The definition of F_1 score is

$$F_1 = \frac{2pr}{p+r}\ , \tag{5.18}$$

where r is recall and p is precision (also called positive predictive value). Standard definitions of recall and precision are

$$r = \frac{c}{c+c'} \text{ and } p = \frac{c}{c+d}\ , \tag{5.19}$$

where c is true positive, c' is the type II error, and d is the type I error. The luggage screening application has a multi-class segmentation problem. Therefore, the standard definition of the precision and recall, given in Eq. 5.19, cannot be used. Our multi-class adaptation defines recall and precision as

$$r = \frac{\sum_{i=1}^{G} N_{G,S}(i, j'(i))}{\sum_{k=1}^{G}\sum_{l=0}^{S} N_{G,S}(k, l)}$$

$$p = \frac{\sum_{i=1}^{G} N_{G,S}(i, j'(i))}{\sum_{k=1}^{G}\sum_{m=1}^{G} N_{G,S}(m, j'(k))}\ .$$

In the above equation $j'(i)$ is the MS label that best matches GT label i as per the Hungarian algorithm matching. Using the equation for precision

given above, we penalize missing portions of segments, missing segments, and split segments equally. The denominator may not include all the MS labels S, because that would penalize splitting more than missed detection (which is unreasonable).

5.4 Synthetic problems

To evaluate our measures against intuitive reasoning, we generated simple problems with different kinds of errors. We consider splitting, merging, partial splitting and merging, and missed objects (type II errors). We do not consider spurious objects (type I errors) because we do not penalize them, as described earlier. The different cases illustrate the behavior of the evaluation measures, including singularities, discontinuities and non-linearities. There are eight cases in which the GT has two object labels, each with 500 voxels. The cases are shown as confusion matrices in Fig. 5.3. In an ideal segmentation, the only populated cells would be along the matrix diagonal. Cases 1-5 consist of errors in which one or more voxels from the first label are misclassified as belonging to the second label as shown below. Considering Case 1, there are two MS labels, but one voxel from segment 1 is mis-classified as belonging to segment 2. This error splits GT segment 1 and merges with GT segment 2. The results of applying the various evaluation measures to this case are in the column labeled Case 1 of Table 5.1. Similarly other columns contain the results for the other cases. In Case 9, one GT label (plus air) is split in two by the MS.

	S_1	S_2
G_1	x	y
G_2	0	500

(a)

	S_1	S_2
G_1	499	1
G_2	1	499

(b)

	S_0	S_1	S_2
G_1	500	0	0
G_2	0	0	500

(c)

	S_1	S_2
G_1	0	500
G_2	0	500

(d)

	S_1	S_2
G_1	500	500

(e)

Figure 5.3: Confusion matrices for the synthetic problems. Cases 1-5 are shown in (a). Case 1: $x = 499$, $y = 1$, Case 2: $x = 475$, $y = 25$, Case 3: $x = 450$, $y = 50$, Case 4: $x = 400$, $y = 100$, Case 5: $x = 250$, $y = 250$. Case 6: one pixel from each GT label is misclassified by MS as belonging to the other label (b), Case 7: One GT label is not detected (c), Case 8: Both GT labels are merged by MS (d), Case 9: Single GT label is split by MS (e).

We compare our measures against OCE and F_1. There are discontinuities in the OCE, but not in the other measures. The OCE jumps from zero at perfect segmentation to 0.25 in our two-label problem when a single pixel is misclassified (Case 1). This is because OCE treats it as a new segment of equal importance as the segment that is a near-perfect match for the GT label. If more voxels are moved over (Cases 2-5), the OCE monotonically increases. However, if instead, one pixel is moved from the second label to the first, as shown by Case 6, there is another jump from 0.25 to 0.5. The discontinuities are an undesirable property of OCE. We contrast Case 6 with Case 2. Intuitively, the error is smaller in Case 6 than 2, and less significant in the luggage screening application, but OCE says the opposite and gives a poorer score to Case 6. In Case 7, there is no penalty for missing an entire object, demonstrating another undesirable property of OCE.

F_1^m monotonically decreases as error increases. It penalizes merging more than missing or split segments, as shown by Cases 7-9, according to the argument that we have not only missed one object but expanded another. This is a combined type I and type II error, which does not occur for two-class problems. However, we could argue that there is only one underlying error, and that we want the merged segments to be penalized no worse than the other types of error. But that is a limitation of the F_1-score definition. Note that the confusion matrix can be weighted, e.g., to assign greater penalty to type II error, although we have not done so here.

The WMI scores are intuitive. While there is a degeneracy to zero for single segments in the GT or MS as shown in Cases 7 -9, we have not encountered this case in our luggage data.

Now we consider the FDR method comprising residuals and slope. The residuals (CVM, R_{L_1}, KL) report the total error. They do not distinguish between Cases 7-9. They give a perfect score of zero for perfect recovery of the feature (volume in these cases), even if the segmentation boundaries are wrong, as illustrated by Case 6. From the point of view of feature recovery it is acceptable to report an error of zero. The slope, shown by K in the table, tells us the kind of error, i.e., splitting or merging in Cases 8 and 9. Case 8 shows undersegmentation; the MS

labels have larger magnitudes than the GT, which occurs if the algorithm merges objects more than it splits them. Conversely, Case 9 has oversegmentation; the MS labels have smaller magnitudes than their corresponding GT labels, which occurs due to splitting. A slope of one tells us that the errors are random (for non-zero residuals). In these examples, no weighting was applied, but can be, for example, to increase the residual penalty for type II errors.

The FDR method is a framework within which we can measure label-wise features. Therefore, we must use residuals that allow different numbers of labels in each image. R_{L_1} gives us a result that is linear with the error, which makes it easy to understand. The CVM errors are monotonic, but nonlinear because CVM accumulates squared errors. The KL divergence is infinite in Cases 7 and 8. The absence of any one label, due to missed objects or merging, results in a division by zero and causes the divergence to be infinite. Therefore we can not use the KL divergence on our luggage data. We have not used another common divergence, the Kolmogorov-Smirnoff divergence (KS2) because it gives undue weight to just one object, which is not desirable in our application, where there are multiple errors and where the range of error magnitude is unpredictable.

In summary, we find that WMI and FDR methods provide acceptable and complementary results for the synthetic problems. We use R_{L_1} because it is linear and permits different numbers of labels in the GT and MS. WMI and FDR are flexible because we can use them on features rather than voxels, and can use weighting as will be demonstrated with bag data.

Table 5.1: Performance values for synthetic test cases considering two GT object labels (and air). The slope of the line fitted to the FRS data is denoted K. CVM, R_{L_1} and KL are the different residual errors. Performance values for perfect MS (no error) are given in the second column as a reference.

	Ideal	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7	Case 8	Case 9
OCE	0	0.25	0.29	0.33	0.39	0.51	0.5	0	0.5	0.5
F_1^m	1	0.999	0.975	0.95	0.9	0.75	0.998	0.67	0.5	0.67
WMI	1	0.99	0.86	0.76	0.62	0.35	0.98	0	0	0
CVM	0	$5\text{x}10^{-4}$	0.0125	0.025	0.05	0.125	0	0.25	0.25	0.5
R_{L_1}	0	10^{-3}	0.025	0.05	0.1	0.25	0	0.5	0.5	0.5
KL	0	≈ 0	0.0013	0.005	0.02	0.144	0	∞	∞	0.693
K	1	1	1	1	1	1	1	1	2	0.5

5.5 Bag data

In this section we present the results of applying our methods to the ALERT luggage images and their segmentations. We first discuss WMI results, then FDR results, and then the human expert validation. In the tables and figures below, we name the segmentation algorithms A1-A5 to anonymize the research groups. The bags are named B1-B5.

5.5.1 WMI results

WMI scores for volume and mass are shown in Tables 5.2 and 5.3 respectively. The tables show that the best performer for volume and mass recovery is algorithm A2. Comparing the mass and volume WMI tables, we see that mass and volume give numerically different results. This happens when the GT or MS labels have a mixture of CT densities. For example, the segmentation algorithms recovered the soles of shoes, not the uppers. In general, the mass scores are higher than the volume scores.

Table 5.2: WMI scores for volume. The best performance in most bags is from A2.

	A1	A2	A3	A4	A5
B1	0.22	0.63	0.54	0.48	0.50
B2	0.45	0.62	0.58	0.48	0.41
B3	0.59	0.69	0.65	0.56	0.38
B4	0.33	0.59	0.65	0.53	0.50
B5	0.60	0.78	0.74	0.68	

Table 5.3: WMI scores for mass. The best performance in most bags is from A2.

	A1	A2	A3	A4	A5
B1	0.27	0.76	0.65	0.58	0.64
B2	0.57	0.74	0.71	0.56	0.58
B3	0.66	0.74	0.69	0.50	0.49
B4	0.40	0.69	0.74	0.63	0.64
B5	0.66	0.84	0.77	0.63	

For comparison, the F_1^m scores for volume and mass are given in Tables 5.4 and 5.5 respectively. The F_1^m scores for volume do not yield a clear winner, but

the scores for mass are similar to the WMI scores in that the mass scores show the best performer to be A2, and the mass scores are generally higher than the volume scores.

Table 5.4: F_1^m scores by volume. It is not clear which is the best performing algorithm.

	A1	A2	A3	A4	A5
B1	0.32	0.67	0.56	0.60	0.55
B2	0.53	0.60	0.57	0.61	0.44
B3	0.49	0.62	0.59	0.57	0.47
B4	0.42	0.52	0.65	0.65	0.59
B5	0.67	0.78	0.76	0.79	

Table 5.5: F_1^m scores by mass. The best performance in most bags is by A2.

	A1	A2	A3	A4	A5
B1	0.37	0.73	0.61	0.68	0.60
B2	0.60	0.68	0.62	0.66	0.56
B3	0.54	0.65	0.61	0.54	0.57
B4	0.46	0.61	0.70	0.70	0.70
B5	0.73	0.83	0.73	0.73	

The WMI scores for uniformity are in Table 5.6. The best performer for the uniformity feature is unclear. Although WMI gave the highest scores to algorithm A2 by volume and mass, A2 is not the best algorithm to recover the uniformity feature.

Table 5.6: WMI score for uniformity. It is not clear which algorithm performs best for this feature.

	A1	A2	A3	A4	A5
B1	0.28	0.77	0.69	0.68	0.66
B2	0.66	0.76	0.75	0.68	0.54
B3	0.68	0.67	0.70	0.67	0.50
B4	0.43	0.71	0.78	0.73	0.64
B5	0.78	0.83	0.87	0.90	

Finally we show the cell-wise WMI weights in Table 5.7. Some WMI scores increase and some decrease compared to unweighted scores, but are not much different from unweighted scores. The results are discussed in more detail in Section 5.6. This weighting does not have a counterpart in the FDR method.

Table 5.7: WMI for cell-wise weighting of volume.

	A1	A2	A3	A4	A5
B1	0.20	0.61	0.51	0.46	0.47
B2	0.43	0.59	0.57	0.46	0.37
B3	0.60	0.69	0.65	0.55	0.36
B4	0.29	0.57	0.64	0.51	0.48
B5	0.59	0.78	0.75	0.67	

5.5.2 FDR results

An example FRS plot is shown in Fig. 5.4 for one algorithm. The FRS slopes for volume, mass and uniformity features, for all algorithms are given in Table 5.8, and the R_{L_1}-residuals are given in Table 5.9 for the combined set of bags. The residual errors per bag for the different features are given in Tables 5.10 through 5.12.

Mass and volume features increase monotonically with the number of voxels in a segment, so a slope $K > 1$ indicates systematic undersegmentation and $K < 1$ indicates systematic oversegmentation, including missing parts of segments. For a non-monotonic feature such as uniformity, FRS slope values do not indicate splitting or merging of the object, but rather a systematic over- or under-estimation of the feature. Over or under-segmentation should not be simplistically defined by counting the number of segments. For example, if multiple machine segments exist for a single GT label, there is oversegmentation. However, if most of the feature is recovered in one machine segment, there is less oversegmentation than if the feature is distributed equally among the multiple machine segments.

Table 5.8: Slopes (K) for FRS fit lines for volume, mass and uniformity features.

	A1	A2	A3	A4	A5
Volume	0.59	0.85	0.56	0.73	0.61
Mass	0.70	1.0	0.58	0.67	0.89
Uniformity	1.26	0.51	0.91	1.06	1.5

Among the algorithms, A2 exhibits best mass and volume recovery. Its FRS slopes are closest to one (Table 5.8), and the residuals are smallest (Table 5.9). For all algorithms, the mass slopes are closer to one than the volume slopes. The mass residuals are also smaller than the volume residuals. As in the WMI scores,

Figure 5.4: The mass scatter plot from algorithm A1. There were 81 GT labels. The fitted line is forced to pass through zero.

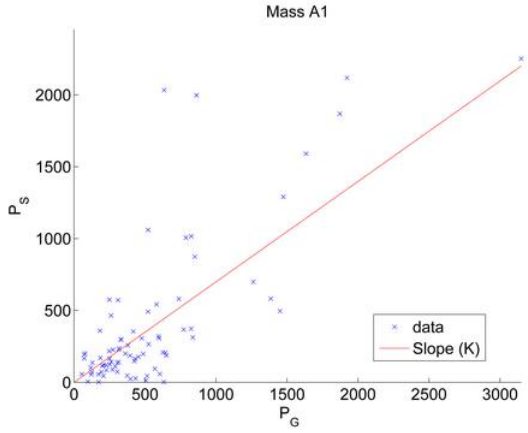


Table 5.9: R_{L_1} residuals for all bags combined. A2 has the smallest residuals.

	A1	A2	A3	A4	A5
Volume	0.49	0.37	0.48	0.44	0.54
Mass	0.41	0.28	0.45	0.44	0.41
Uniformity	0.60	0.33	0.54	0.51	0.62

the FRS plots show that it is easier for a segmentation algorithm to recover mass than volume because of the heterogeneity of the material composition of objects and clutter.

Although A2 has the best volume and mass retrieval, it does not show best recovery of uniformity. As shown in Table 5.8, the uniformity slope for A2 is small (0.51) compared to other algorithms. There is also no clear best performer. The FDR results are in line with the WMI scores.

An instance of poor uniformity recovery was a water bottle touching another liquid-filled container. The MS label for the water bottle included the other liquid, and lost some of the bottle itself, either labeling it as air or as the other liquid as shown in Fig 5.5. Also included into the bottle label were voxels of metal from a nearby touching object (not shown). The volume and mass were well-recovered because of the exchange of material between the two labels, but the CT number differences of the mixed materials created a high variance in the machine segment.

Feature recovery over the feature range is shown in Fig 5.6 for mass. The sliding average scatter plots show that feature recovery improves as object mass

Table 5.10: R_{L_1} residual error by volume. A2 has the smallest residual in most bags.

	A1	A2	A3	A4	A5
B1	0.76	0.46	0.61	0.56	0.51
B2	0.45	0.51	0.58	0.51	0.59
B3	0.37	0.27	0.43	0.48	0.60
B4	0.66	0.44	0.50	0.49	0.45
B5	0.37	0.22	0.36	0.30	

Table 5.11: R_{L_1} residual error by mass. A2 has the smallest residual in most bags.

	A1	A2	A3	A4	A5
B1	0.71	0.35	0.55	0.48	0.41
B2	0.38	0.41	0.53	0.46	0.44
B3	0.32	0.23	0.39	0.53	0.48
B4	0.63	0.35	0.43	0.43	0.31
B5	0.31	0.15	0.40	0.38	

increases for A2. Although A2 has the best mass scores, it is less reliable for low-mass objects than some of the other algorithms. At higher masses, it is more reliable than the other algorithms. In another example, A5 shows no apparent preference for any range of mass. An example plot of outliers is shown in Fig. 5.7 for the mass feature, for the A1 algorithm.

5.5.3 Validation by human expert observer

In order to validate the methods beyond the synthetic problems, a human expert visually evaluated two MS algorithms, A1 and A2, against GT. The difficulty of the task for the observer arises from the multiple splits and merges and the large number of slices. The comparison was simplified by sampling every fifteenth slice. The observer was presented with corresponding slices of A1, A2, GT and CT images. The MS slices were randomly ordered for blind review. For each pair of slices, the observer selects the MS that he considers a closer match to the GT slice. The results are in Table 5.13. In each bag, the observer prefers A2 over A1, which is in agreement with WMI and FDR results. The expert explained some of his decisions. He observed that slices from A1 had more type II error than those

Table 5.12: R_{L_1} residual error by uniformity. A2 has the smallest residuals despite the small slope shown in Table 5.8.

	A1	A2	A3	A4	A5
B1	1.18	0.39	0.71	0.55	0.40
B2	1.00	0.35	1.01	0.97	0.90
B3	0.92	0.24	0.64	0.51	0.42
B4	0.41	0.33	1.03	0.78	0.62
B5	0.42	0.35	0.30	0.35	

Table 5.13: The human observer evaluation of two MS algorithms. The second column shows percentage of slices in which A2 was preferred. The third column shows the number of slices that were ranked better in A1, in A2 and equal in both. The fourth column shows the percentage of slices with a higher WMI score when WMI was applied slicewise.

Bag	**% A2 by Expert**	**A1 / A2/ Equal**	**% A2 by MI**
B1	96	0 / 27 / 1	100
B2	82	5 / 32 / 2	85
B3	73	6 / 22 / 2	91
B4	76	6 / 25 / 2	86
B5	72	8 / 26 / 2	92

from A2. This observation relates to the lower WMI scores (Table 5.2 and 5.3) and smaller slope of A1 (Table 5.8) compared with A2. The expert selected A1 in some slices where A2 labels appeared jagged. The jagged labels belonged to large liquid-filled containers. These selections agree with the uniformity scores.

There is also a correspondence between the expert's preferred percentage and the WMI and R_{L_1} results per bag. We do not expect to see perfect correspondence because the expert performed a simplifed evaluation. The slice sampling method favors larger less dense objects over smaller denser ones, the human is imprecise and is influenced by visual appeal, there was no weighting per slice to increase the impact of fuller slices over emptier ones, and no quantification of preference given a pair of slices.

In addition, we applied WMI slicewise on the same slices evaluated by the human. For all slices, the higher-scoring algorithm was compared with the human preference using McNemar [108] and KS2 tests. The McNemar test yielded a p-value of 0.08 which does not reject the null hypothesis that the human and WMI prefer the same algorithm. The KS2 test-statistic was 0.04, which also does not

reject the null hypothesis at a confidence level of 0.05.

5.5.4 Summary

The FDR and WMI both measure feature recovery, unlike existing evaluation methods that compute edge distances or voxel misclassification. Both methods are sensitive to spatial correspondence of labels, unlike histogram comparison methods that measure features. Further, both methods are useful for multiple label segmentation problems. And both allow us to assign priorities to segments by pointwise or regional weighting. They also gave consistent results in selecting the same best algorithm. However, FDR and WMI have different perspectives. WMI is more sensitive to spatial correspondence than FDR. FDR is more flexible in that data from multiple images can be pooled and trends can be extracted, and a wider variety of features can be used. A human expert validated our methods by visual assessment.

5.6 Discussion

As discussed in Section 5.1, many GT-based methods in the evaluation literature use region-based errors when multiple regions of interest are present in the image. This can be thought of as using an indicator function on each voxel for each label. But each voxel and its neighborhood contain additional information we can use instead of just the indicator. In our case, we have used mass and uniformity in addition to volume. In the mass scores, voxels with higher CT number are more important than those with lower CT number. The use of uniformity prioritizes more homogenous objects over less homogenous ones. Mass and uniformity are examples of features that may be useful for a specific application. An EDS may utilize these or other features depending on the ATR algorithm.

The WMI, F_1^m score and FDR results for mass are more consistent with each other (same best performer) than the corresponding volume scores. These discrepancies between volume and mass illustrate the challenges of segmentation of CT images of luggage. The results show that mass is easier to recover than volume,

i.e., a meaningful feature within a region is easier to extract than the region itself.

The FDR method is more general and informative than histogram-based methods. A previously published evaluation method for populations of similar objects used histograms [12]. However, in general, the objects in a segmentation problem are not similar, and there are no object-type populations. We generate a bipartite matching and can evaluate any objects. Due the bipartite matching, we can extract information about systematic errors, expected performance as a function of feature value, and outliers. Matching allows object prioritization and non-uniform costs. The residuals include pairwise errors, missed and spurious segments (although we do not penalize the latter here).

The FRS and WMI showed that A2 traded-off region uniformity for better overall segmentation. Note that if the CT number distribution of adjoining objects is the same, the mass or volume FRS plots may not indicate errors (provided the same volumes are displaced from one object to another). If the textures are similar as well, the FRS plot for uniformity will not indicate errors either. This is acceptable from the feature recovery point of view. Another inference we can draw from the uniformity results is that the improvement of the other algorithms relative to A2 shows that they find it easier to segment uniform objects, while A2 is less dependent on object uniformity.

Our sliding-average plots show trends in performance as a function of feature value for some algorithms. We show the mass feature, because that has the best WMI and FDR scores. The accuracy of segmentation of an object depends not only on its own features, but those of the surrounding objects. As a result, algorithms may not all show trends with feature value, but if trends are present, they help in the interpretation of segmentation results.

In the cell-wise weighted confusion matrix, we have weighted each cell by a factor representing the similarity of a regional feature. Our factor is the ratio of the smaller mean to the greater mean. For a cell representing some GT and MS labels, if the labels are dissimilar in the regional feature, we assign a smaller weight to the cell, which is to say that this cell does not help us get information about one distribution from the other distribution. This decreases the total WMI.

Consider a cell on the diagonal of the confusion matrix. The diagonal represents the matched objects. If the matched objects are dissimilar, then the ratio is small, and the cell loses importance. Here it is easy to see the interpretation that the object represented by the cell in the MS image does not tell us much about the GT image. Considering an off-diagonal cell, it similarly loses importance when the means are dissimilar. At first glance, it seems counter-intuitive that a cell that represents two unmatched objects, should have decreased weight when the objects are dissimilar. But WMI does not measure the ordering of the information. This cell contains the quantity of an intersection that really does exist. So if we decrease (increase) the weighting of that intersection, we decrease (increase) the amount of information one label set tells us about the other label set. In addition, we increase or decrease the entropies of the GT and MS images when we weight cells, depending on what the original image contained. The cell-wise weighting therefore is difficult to control and does not give monotonic results.

5.7 Evaluation of metal artifact reduction

We have used our segmentation evaluation algorithms to evaluate the effectiveness of our metal artifact reduction (MAR) algorithm. Although we have a more rigorous and application-independent numerical evaluation of the MAR algorithms in Chapter 3, evaluation of the MAR by segmentation results allows us to directly measure the impact of MAR on the application. For this evaluation, we choose region growing to segment the original and MAR images. This is because we are evaluating the reconstruction of homogenous objects, which is what region growing is good at segmenting. We show the results from four images in our set. These images have multiple objects of interest, so that we do not have a degenerate case for WMI.

The GT for these images was manually generated for the numerical evaluation of Chapter 3. Since we were measuring uniformity, we placed the manual contours within the object, excluding the high-contrast edges. For this evaluation by segmentation, we dilated each GT label by two pixels so that the label would

Table 5.14: Volume WMI for the four images that contain multiple uniform objects

	Bag 1	**Bag 3**	**Bag6**	**Bag 7**
Original	0.87	0.70	0.69	0.71
MAR	0.94	0.80	0.87	0.95

extend to the edge of the object, and be close to an ideal segmentation. The CT images and the MS labels for original and MAR images are shown in Fig. 5.8.

The WMI results for volume are given in Table 5.14. The table shows that the WMI scores improve after MAR. Fig. 5.9 shows the FRS plots for the original and MAR images. The FRS slopes show that oversegmentation has decreased. Dark streaks split the uniform objects in the original images, creating smaller fragments, but after MAR, more of the objects are segmented correctly. The R_{L_1} residual error for the original image set is 0.22 while the error for the MAR set is 0.13. From each of these results, we see that MAR improves the image segmentation.

5.7.1 Conclusion

We have developed two flexible parameter-independent methods to evaluate segmentation algorithms. The methods were applied on a test set of luggage images. Our contributions are as follows. 1. We have used a well-accepted measure from information theory to measure feature overlap. 2. We have developed a new method based on feature recovery that has good agreement with mutual information, but that also identifies systematic errors and allows pointwise or regional features to be used. 3. We have used weighting functions to prioritize objects based on desired features. 4. We developed a semi-automatic method to extract GT from three-dimensional CT images. We used human evaluation of segmentation accuracy and synthetic problems to validate our methods. Our evaluation methods indicated one algorithm, A2, as the best one, and found characteristics of the algorithm: accuracy increased with object mass, and A2 was less reliant on object uniformity than some of the other algorithms. Given the challenges and requirements for segmentation in luggage scanning, we found our methods to be

more suitable to evaluate segmentation algorithms than methods from existing literature.

In addition, we were able to use this to help evaluate our metal artifact reduction. Both evaluation methods indicated that each MAR image was superior to the corresponding original image with artifacts, which is in line with our visual and quantitative evaluation in Chapter 3.

This chapter contains material from "Flexible Methods for Segmentation Evaluation: Results from CT Luggage Screening," *Journal of X-Ray Science and Technology*, Vol. 22, Issue 2, 2014. The paper was co-authored by Xiaoqian Jiang, Pamela Cosman and Harry Martz.

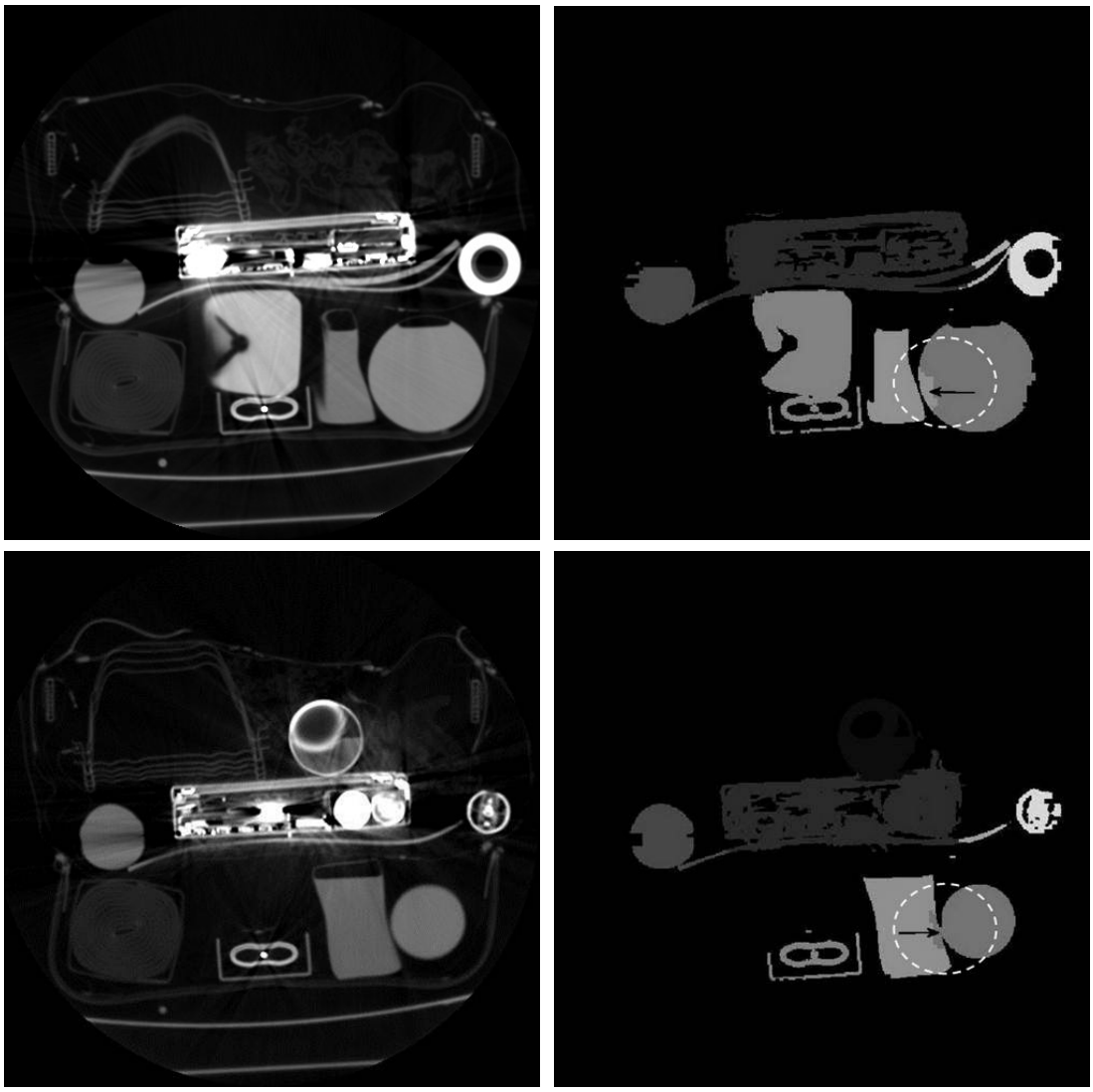


Figure 5.5: Poor uniformity recovery by A2 of a large uniform object. Two objects circled in the right column are liquid-filled containers. There is misclassification between those two object labels, shown by arrows, as well as one of the objects and air.

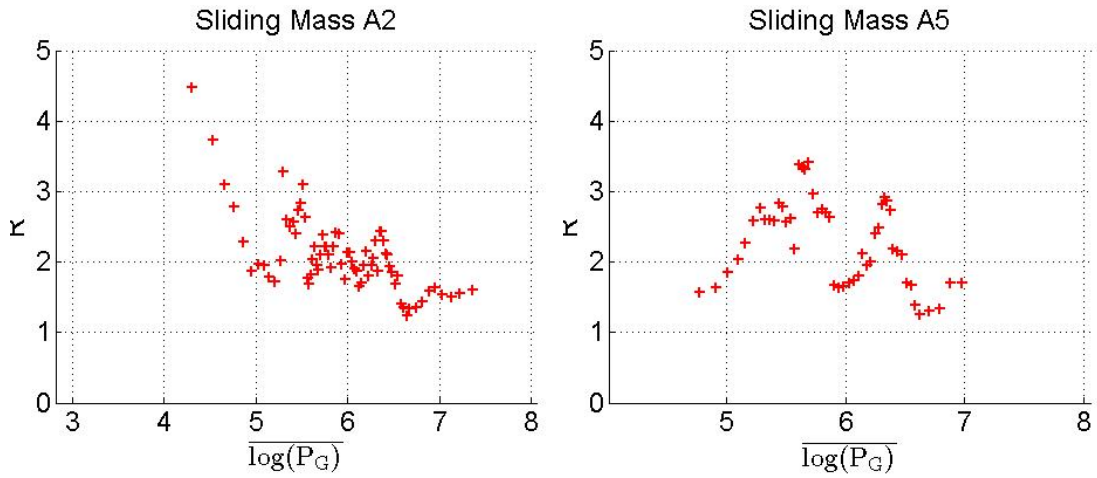


Figure 5.6: The sliding average (Eq. 5.16) for the mass feature shown for two algorithms have different characteristics. A2 improves with mass, but A5 does not.

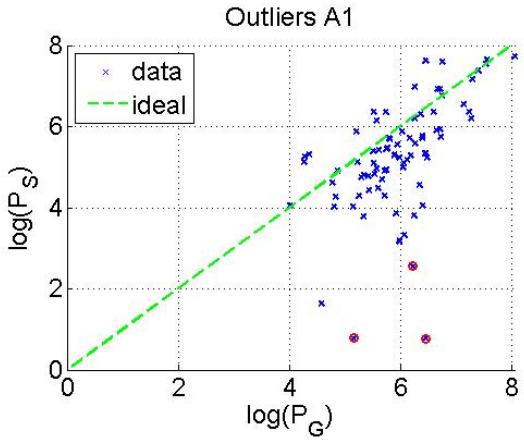


Figure 5.7: Example FRS plot for mass showing outliers (Eq. 5.17) circled in red.

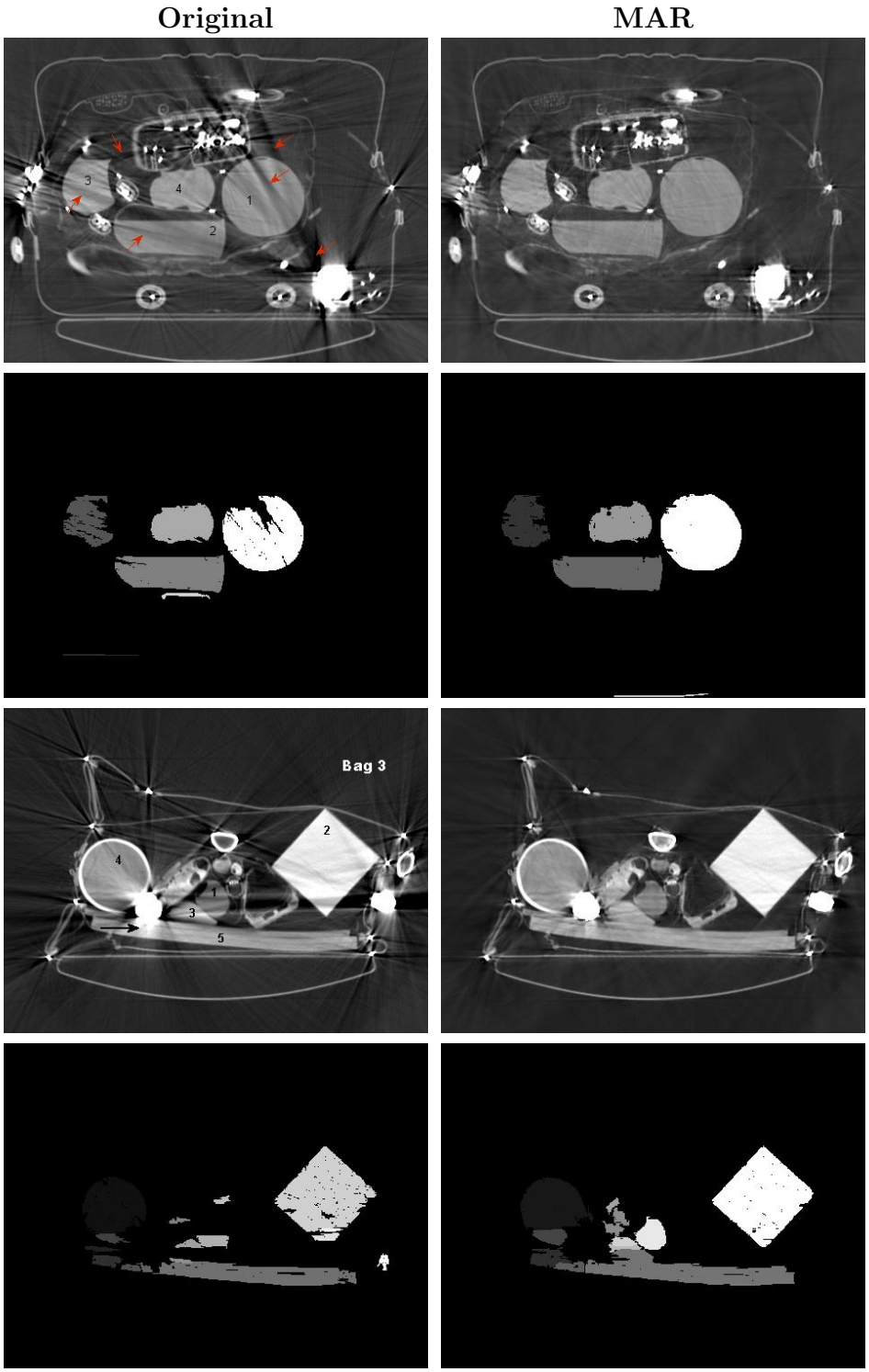


Figure 5.8: CT images and their segmentations.

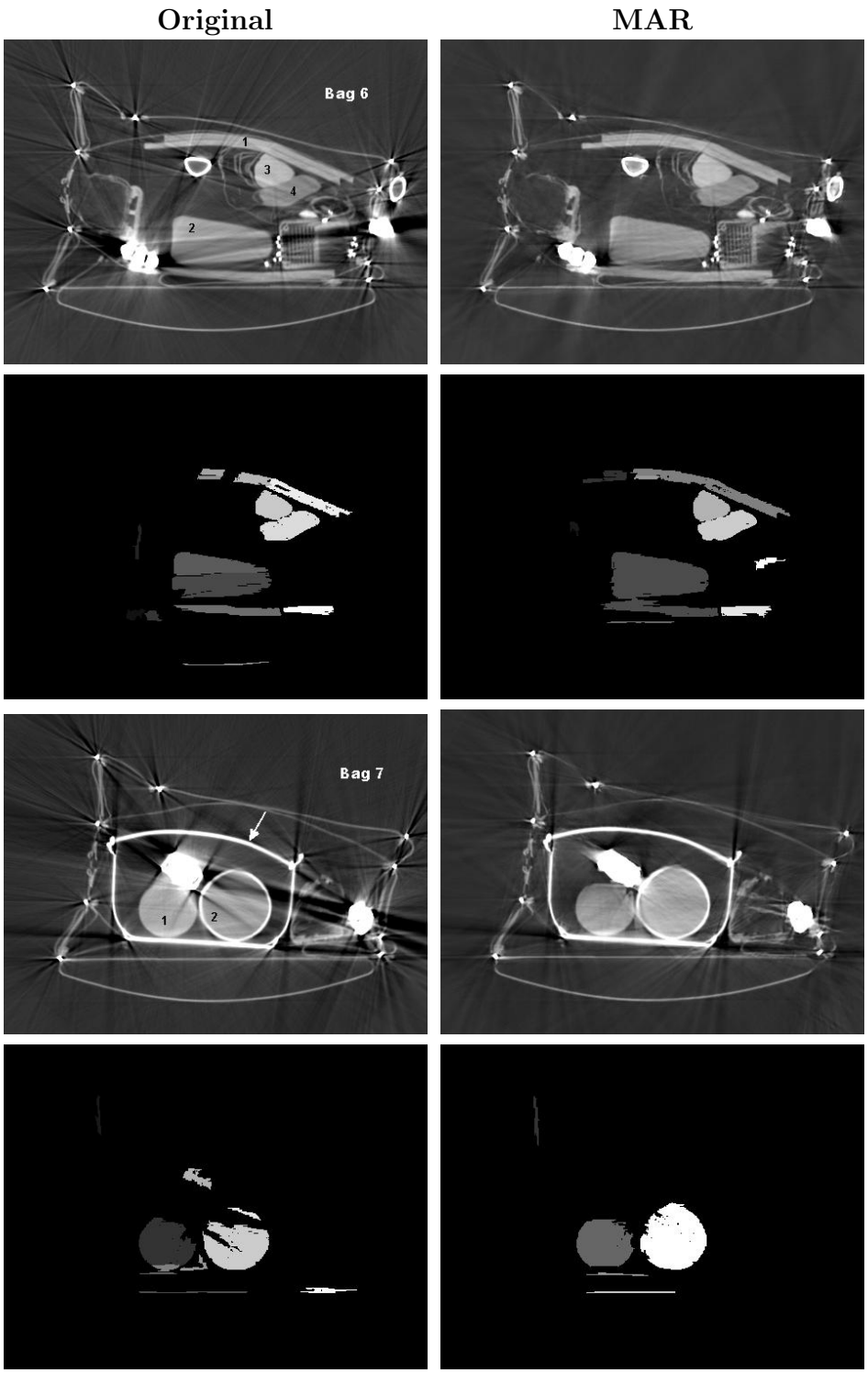


Figure 5.8: CT images and their segmentations (continued).

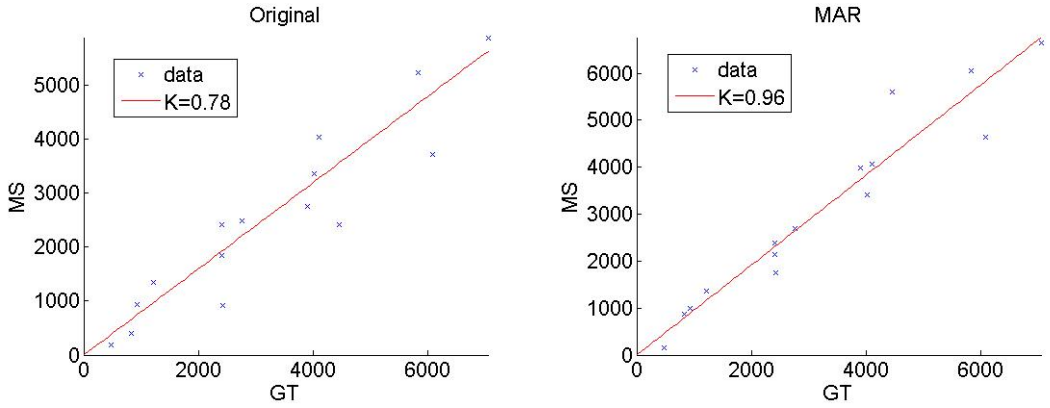


Figure 5.9: FRS plots for the original and MAR image sets. The slope K of the original set indicates splitting by metal artifacts. The slope of the MAR image set is closer to the ideal value of 1.

Appendix A

Ancillary operations for segmentation-based MAR algorithm

A.1 Contouring the outer boundary

We create a closed contour along the outer boundary of the anatomy in the original image, shown by the broken line in Fig. 2.1(a). The contour is assigned a low value of soft tissue such as $I_t = -100$ HU. The contour must be at least two voxels thick to prevent the later steps from removing it. There are many contour tracking algorithms which would work for this simple contour generation problem; we have used an algorithm that thresholds the image, and then connects outer boundary pixels with 8-connectivity. We use a threshold of -600 HU, which is well below soft-tissue intensities, to account for tissue corrupted by artifacts. If there are multiple disconnected objects in the image, and hence multiple contours, we keep only the largest one.

A.2 Correcting errors from morphology approximations

OBR and CBR replace regions of voxels with single values. If left this way, the prior-image would be patchy, and the final image would have the appearance of patchy texture. To avoid this, voxel values between the limits of I_t and I_{min} are replaced with the mode value of the original image. This range includes most soft tissue, but the exact limits are not critical, and the range can be made larger. Soft tissue variations will be removed from the prior. The soft tissue variations do not contribute to secondary artifacts, and it is better to replace them, to avoid patchiness from CBR and OBR. The data replacement method, described below, does not substitute the scanner projections with the reprojections of the prior, so the soft tissue details are not lost.

A.3 Replacement of sinogram data

The prior-image is reprojected, and the metal trace is found in the reprojections by calculation of the rays passing through the metal image. We have used the method described in [15] for data replacement. In this method, the reprojections are subtracted from the scanner projections. The difference projections are smoother than the scanner projections. The interpolation of metal traces is done on the difference projections and the interpolated result is added to the reprojections to create the final corrected projections that are then reconstructed to create the final image. We modify this interpolation method in that we fit a second order spline to five samples on both sides of the metal trace instead of using linear interpolation. Linear interpolation of two samples should not be relied upon because sampling errors and noise will result in poor estimates of data.

Bibliography

[1] "Final report on algorithm development for security applications," in *Algorithm Development for Security Applications*, M. Silevitch, C. Crawford, and H. Martz, Eds., Boston, 2009.

[2] A. Kak and M. Slaney, *Principles of computerized tomographic imaging*. Philadelphia, PA 19104: SIAM, 1988.

[3] F. Natterer, *The Mathematics of Computerized Tomography*, 1st ed. Chichester: Wiley and Sons, 1986.

[4] J. Hsieh, *Computed tomography: Principles, Design, Artifacts, and Recent Advances*, 2nd ed. SPIE–The International Society for Optical Engineering, 2009.

[5] K. Lange and R. Carson, "EM reconstruction algorithms for emission and transmission tomography," *Journal of Computer Assisted Tomography*, vol. 8, no. 2, pp. 306–316, 1984.

[6] J.-B. Thibault, K. D. Sauer, C. A. Bouman, and J. Hsieh, "A three-dimensional statistical approach to improved image quality for multislice helical CT," *Medical Physics*, vol. 34, no. 11, p. 4526, Oct. 2007.

[7] R. Gordon, R. Bender, and G. Herman, "Algebraic reconstruction techniques (ART) for three-dimensional electron microscopy and X-ray photography," *Journal of Theoretical Biology*, vol. 29, pp. 471–481, 1970.

[8] B. D. Man and J. Nuyts, "Metal streak artifacts in X-ray computed tomography: a simulation study," in *Nuclear Science Symposium*, 1998, pp. 1860–1865.

[9] E. de Paula, "XSPEC," 1984.

[10] M. Berger and J. Hubbell, "XCOM, photon cross sections on a personal computer," Gaithersburg, MD, 1987.

[11] P. Joseph and R. Spital, "A method for correcting bone induced artifacts in computed tomography scanners," *Journal of Computer Assisted Tomography*, vol. 2, no. 1, pp. 100–108, 1978.

[12] G. Glover and N. Pelc, "An algorithm for the reduction of metal clip artifacts in CT reconstructions," *Medical Physics*, vol. 8, pp. 799–807, 1981.

[13] W. Kalender, R. Hebel, and J. Ebersberger, "Reduction of CT artifacts caused by metallic implants," *Radiology*, vol. 164, no. 2, pp. 576–77, 1987.

[14] S. Zhao, D. D. Robertson, G. Wang, B. Whiting, and K. T. Bae, "X-ray CT metal artifact reduction using wavelets: an application for imaging total hip prostheses." *IEEE Transactions on Medical Imaging*, vol. 19, no. 12, pp. 1238–47, Dec. 2000.

[15] R. Naidu, I. Bechwati, S. Karimi, S. Simanovsky, and C. Crawford, "Method of and system for reducing metal artifacts in images generated by x-ray scanning devices," U.S. Patent 6 721 387, 2004.

[16] M. Bal and L. Spies, "Metal artifact reduction in CT using tissue-class modeling and adaptive prefiltering," *Medical Physics*, vol. 33, no. 8, pp. 2852–2859, Jul. 2006.

[17] C. Golden, S. R. Mazin, F. E. Boas, G. Tye, P. Ghanouni, G. Gold, M. Sofilos, and N. J. Pelc, "A comparison of four algorithms for metal artifact reduction in CT imaging," in *SPIE Medical Imaging*, N. J. Pelc, E. Samei, and R. M. Nishikawa, Eds. International Society for Optics and Photonics, Mar. 2011, p. 79612Y.

[18] E. Meyer, R. Raupach, M. Lell, B. Schmidt, and M. Kachelrieß, "Frequency split metal artifact reduction (FSMAR) in computed tomography." *Medical Physics*, vol. 39, no. 4, pp. 1904–16, Apr. 2012.

[19] F. E. Boas and D. Fleischmann, "Evaluation of two iterative techniques for reducing metal artifacts in computed tomography." *Radiology*, vol. 259, no. 3, pp. 894–902, Jun. 2011.

[20] J. M. Verburg and J. Seco, "CT metal artifact reduction method correcting for beam hardening and missing projections." *Physics in Medicine and Biology*, vol. 57, no. 9, pp. 2803–18, May 2012.

[21] T. Koehler, B. Brendel, and K. Brown, "A New Method for Metal Artifact Reduction in CT," in *The International Conference on Image Formation in X-ray Computed Tomography*, Salt Lake City, Utah, USA, 2011.

[22] R. Alvarez and A. Macovski, "Energy-selective reconstructions in x-ray computerised tomography," *Physics in Medicine and Biology*, vol. 21, no. 5, pp. 733–744, 1976.

[23] B. De Man, J. Nuyts, P. Dupont, G. Marchal, and P. Suetens, "An iterative maximum-likelihood polychromatic algorithm for CT." *IEEE Transactions on Medical Imaging*, vol. 20, no. 10, pp. 999–1008, Oct. 2001.

[24] F. Bamberg, A. Dierks, and K. Nikolaou, "Metal artifact reduction by dual energy computed tomography using monoenergetic extrapolation," *European Radiology*, vol. 21, no. 7, pp. 1424–1429, 2011.

[25] P. Sukovic and N. Clinthorne, "Penalized weighted least-squares image reconstruction for dual energy X-ray transmission tomography," *IEEE Transactions on Medical Imaging*, vol. 19, no. 11, pp. 1075 – 1081, 2000.

[26] H. Xue, L. Zhang, and Y. Xiao, "Metal artifact reduction in dual energy CT by sinogram segmentation based on active contour model and TV inpainting," in *Nuclear Science Symposium Conference Record (NSS/MIC)*, Orlando, FL, 2009, pp. 904–908.

[27] H. Li, L. Yu, X. Liu, J. G. Fletcher, and C. H. McCollough, "Metal artifact suppression from reformatted projections in multislice helical CT using dual-front active contours," *Medical Physics*, vol. 37, no. 10, p. 5155, Sep. 2010.

[28] C. Zhou, Y. Zhao, S. Luo, H. Shi, and L. Zheng, "Monoenergetic imaging of dual-energy CT reduces artifacts from implanted metal orthopedic devices in patients with factures," *Academic Radiology*, vol. 18, no. 10, pp. 1252–1257, 2011.

[29] G. Wang and D. Snyder, "Iterative deblurring for CT metal artifact reduction," *IEEE Transactions on Medical Imaging*, vol. 15, no. 5, pp. 657–664, 1996.

[30] M. Oehler and T. Buzug, "Modified MLEM algorithm for artifact suppression in CT," in *Nuclear Science Symposium Conference Record*, vol. 6, San Diego, CA, 2006, pp. 3511–3518.

[31] R. Murphy, S. Yan, J. O'Sullivan, D. Snyder, B. Whiting, D. Politte, G. Lasio, and J. Williamson, "Pose estimation of known objects during transmission tomographic image reconstruction," *IEEE Transactions on Medical Imaging*, vol. 25, no. 10, pp. 1392–1404, Oct. 2006.

[32] 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*, vol. 28, no. 2, pp. 250–260, 2009.

[33] J. Stayman and Y. Otake, "Model-Based Tomographic Reconstruction of Objects Containing Known Components," *IEEE Transactions on Medical Imaging*, vol. 31, no. 10, pp. 1837–1848, 2012.

[34] T. Sakimoto and K. Nishino, "Metal artifact reduction in tomosynthesis by metal extraction and ordered subset-expectation maximization (OS-EM) reconstruction," in *SPIE Medical Imaging*, 2013, p. 86685M.

[35] J. Rinkel, W. Dillon, and T. Funk, "Computed tomographic metal artifact reduction for the detection and quantitation of small features near large metallic implants: a comparison of published methods," *Journal of Computer Assisted Tomography*, vol. 32, no. 4, pp. 621–629, 2008.

[36] A. Mouton and N. Megherbi, "An experimental survey of metal artefact reduction in computed tomography," *Journal of X-ray Science and Technology*, vol. 21, no. 2, pp. 193–226, 2013.

[37] S. Mazin and N. Pelc, "SU-EE-A4-03: Metal Artifact Reduction Algorithm for X-Ray CT Using a Three-Pass Approach," *Medical Physics*, vol. 36, no. 6, pp. 2342–2342, 2009.

[38] E. Meyer, R. Raupach, M. Lell, B. Schmidt, and M. Kachelriess, "Normalized metal artifact reduction (NMAR) in computed tomography." *Medical Physics*, vol. 37, no. 10, pp. 5482–93, Oct. 2010.

[39] M. Stille, B. Kratz, J. Müller, N. Maass, I. Schasiepen, M. Elter, I. Weyers, and T. M. Buzug, "Influence of metal segmentation on the quality of metal artifact reduction methods," in *SPIE Medical Imaging*, R. M. Nishikawa and B. R. Whiting, Eds. International Society for Optics and Photonics, Mar. 2013, pp. 86 683C–86 683C–6.

[40] W. J. H. Veldkamp, R. M. S. Joemai, A. J. van der Molen, and J. Geleijns, "Development and validation of segmentation and interpolation techniques in sinograms for metal artifact suppression in CT." *Medical Physics*, vol. 37, no. 2, pp. 620–8, Feb. 2010.

[41] H. Tuy, "A post-processing algorithm to reduce metallic clip artifacts in CT images," *European Radiology*, vol. 3, no. 2, pp. 129–134, 1993.

[42] J. Wang, S. Wang, Y. Chen, J. Wu, J.-L. Coatrieux, and L. Luo, "Metal artifact reduction in CT using fusion based prior image." *Medical physics*, vol. 40, no. 8, p. 081903, Aug. 2013.

[43] L. Grady, V. Singh, T. Kohlberger, C. Alvino, and C. Bahlmann, "Automatic Segmentation of Unknown Objects, with Application to Baggage Security," in *European Conference on Computer Vision*. Florence, Italy: Springer-Verlag, 2012, pp. 430–444.

[44] M. R. Paudel, M. Mackenzie, B. G. Fallone, and S. Rathee, "Evaluation of normalized metal artifact reduction (NMAR) in kVCT using MVCT prior images for radiotherapy treatment planning," *Medical Physics*, vol. 40, no. 8, p. 081701, Jul. 2013.

[45] I. A. Elbakri and J. A. Fessler, "Statistical image reconstruction for polyenergetic X-ray computed tomography." *IEEE Transactions on Medical Imaging*, vol. 21, no. 2, pp. 89–99, Feb. 2002.

[46] ——, "Segmentation-free statistical image reconstruction for polyenergetic x-ray computed tomography with experimental validation," *Physics in Medicine and Biology*, vol. 48, no. 15, pp. 2453–2477, Aug. 2003.

[47] X. Zhang, J. Wang, and L. Xing, "Metal artifact reduction in x-ray computed tomography (CT) by constrained optimization," *Medical Physics*, vol. 38, no. 2, p. 701, 2011.

[48] Y. Zhang, H. Yan, X. Jia, J. Yang, S. Jiang, and X. Mou, "A hybrid metal artifact reduction algorithm for x-ray CT," *Medical Physics*, vol. 40, p. 041910, 2013.

[49] E. Y. Sidky and X. Pan, "Image reconstruction in circular cone-beam computed tomography by constrained, total-variation minimization." *Physics in Medicine and Biology*, vol. 53, no. 17, pp. 4777–807, Sep. 2008.

[50] Y. Nesterov, "Smooth minimization of non-smooth functions," *Mathematical Programming*, vol. 103, no. 1, pp. 127–152, 2005.

[51] S. Becker, J. Bobin, and E. J. Candes, "Nesta: a fast and accurate first-order method for sparse recovery," *SIAM Journal on Imaging Sciences*, vol. 4, no. 1, pp. 1–39, 2011.

[52] K. Haris, S. N. Efstratiadis, N. Maglaveras, and A. K. Katsaggelos, "Hybrid image segmentation using watersheds and fast region merging." *IEEE Transactions on Image Processing*, vol. 7, no. 12, pp. 1684–99, Jan. 1998.

[53] "Final report on Algorithm Development for Security Applications 6," in *Algorithm Development for Security Applications*, M. Silevitch, C. Crawford, and H. Martz, Eds. Boston: Northeastern University, 2011.

[54] R. Gonzalez, R. Woods, and S. Eddins, *Digital image processing using MATLAB*, 1st ed. Pearson, 2009.

[55] L. A. Shepp and Y. Vardi, "Maximum likelihood reconstruction for emission tomography." *IEEE Transactions on Medical Imaging*, vol. 1, no. 2, pp. 113–22, Jan. 1982.

[56] J. Fessler, "Penalized weighted least-squares image reconstruction for positron emission tomography," *IEEE Transactions on Medical Imaging*, vol. 13, no. 2, pp. 290–300, 1994.

[57] G.-H. Chen, J. Tang, and S. Leng, "Prior image constrained compressed sensing (PICCS): A method to accurately reconstruct dynamic CT images from highly undersampled projection data sets," *Medical Physics*, vol. 35, no. 2, p. 660, Jan. 2008.

[58] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.

[59] A. Papoulis and S. U. Pillai, *Probability, Random Variables and Stochastic Processes*, 4th ed. McGraw-Hill Europe, 2002.

[60] J. Dobbs and H. Weedon, "Normalization of tomographic image data," U.S. Patent 5 680 427, 1996.

[61] E. Anderson and K. Anderson, "The MOSEK interior point optimizer for linear programming: an implementation of the homogeneous algorithm," 2013.

[62] P. Jin, S. J. Kisner, T. Frese, and C. A. Bouman, "Model-based iteartive reconstruction (MBIR) software for X-ray CT," Nov 2013, available from http://engineering.purdue.edu/˜bouman.

[63] C. Crawford and S. Song, "Scanning Requirement Specification for the ALERT Reconstruction Initiative (Task Order 3)," Awareness and Localization of Explosives-Related Threats (ALERT), Tech. Rep., 2013.

[64] "Research and Development of Reconstruction Advances in CT-based Object Detection Systems," in *New Methods for Explosive Detection for Aviation Security*, C. Crawford, H. Martz, and M. Silevitch, Eds. Boston: Awareness and Localization of Explosives-Related Threats (ALERT), 2013, pp. 390–397.

[65] J. F. Williamson, B. R. Whiting, J. Benac, R. J. Murphy, G. J. Blaine, J. A. OSullivan, D. G. Politte, and D. L. Snyder, "Prospects for quantitative computed tomography imaging in the presence of foreign metal bodies using statistical image reconstruction," *Medical Physics*, vol. 29, no. 10, p. 2404, Sep. 2002.

[66] B. Kratz, S. Ens, J. Muller, and T. M. Buzug, "Reference-free ground truth metric for metal artifact evaluation in CT images," *Medical Physics*, vol. 38, no. 7, p. 4321, Jun. 2011.

[67] F. O'Sullivan, Y. Pawitan, and D. Haynor, "Reducing negativity artifacts in emission tomography: post-processing filtered backprojection solutions," University of Washington, Tech. Rep., 1993.

[68] Y. Takahashi, S. Mori, T. Kozuka, K. Gomi, T. Nose, T. Tahara, M. Oguchi, and T. Yamashita, "Preliminary study of correction of original metal artifacts due to I-125 seeds in postimplant dosimetry for prostate permanent implant brachytherapy," *Radiation Medicine*, vol. 24, no. 2, pp. 133–138, Feb. 2006.

[69] S. Karimi, P. Cosman, C. Wald, and H. Martz, "Segmentation of artifacts and anatomy in CT metal artifact reduction," *Medical Physics*, vol. 39, pp. 5857–5868, 2012.

[70] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[71] R. Chartrand and W. Yin, "Iteratively reweighted algorithms for compressive sensing," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008.

[72] H. Zou and T. Hastie, "Regularization and variable selection via the Elastic Net," *Journal of the Royal Statistical Society, Series B*, vol. 67, no. 2, pp. 301–320, 2005.

[73] S. Boyd, N. Parikh, and E. Chu, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[74] C. Li, "An efficient algorithm for total variation regularization with applications to the single pixel camera and compressive sensing," Ph.D. dissertation, Rice University, 2009.

[75] Y. J. Zhang, "A survey on evaluation methods for image segmentation," *Pattern Recognition*, vol. 29, no. 8, pp. 1335–1346, 1996.

[76] H. Zhang, J. E. Fritts, and S. A. Goldman, "Image segmentation evaluation: A survey of unsupervised methods," *Computer Vision and Image Understanding*, vol. 110, no. 2, pp. 260–280, May 2008.

[77] B. Johnson and Z. Xie, "Unsupervised image segmentation evaluation and refinement using a multi-scale approach," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 66, no. 4, pp. 473–483, Jul. 2011.

[78] A. Fenster and B. Chiu, in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society.* IEEE Engineering in Medicine and Biology Society, Jan 2005.

[79] T. Heimann, B. van Ginneken, M. a. Styner, Y. Arzhaeva, V. Aurich, C. Bauer, A. Beck, C. Becker, R. Beichel, G. Bekes, F. Bello, G. Binnig, H. Bischof, A. Bornik, P. M. M. Cashman, Y. Chi, A. Cordova, B. M. Dawant, M. Fidrich, J. D. Furst, D. Furukawa, L. Grenacher, J. Hornegger, D. Kainmüller, R. I. Kitney, H. Kobatake, H. Lamecker, T. Lange, J. Lee, B. Lennon, R. Li, S. Li, H.-P. Meinzer, G. Nemeth, D. S. Raicu, A.-M. Rau, E. M. van Rikxoort, M. Rousson, L. Rusko, K. a. Saddi, G. Schmidt, D. Seghers, A. Shimizu, P. Slagmolen, E. Sorantin, G. Soza, R. Susomboon, J. M. Waite, A. Wimmer, and I. Wolf, "Comparison and evaluation of methods for liver segmentation from CT datasets." *IEEE Transactions on Medical Imaging*, vol. 28, no. 8, pp. 1251–65, Aug. 2009.

[80] Y. Chang, J. Xia, P. Yuan, and T. Kuo, "3D segmentation of maxilla in cone-beam computed tomography imaging using base invariant wavelet active shape model on customized two-manifold topology," *Journal of X-ray Science and Technology*, vol. 21, pp. 251–282, 2013.

[81] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, vol. 2. IEEE Comput. Soc, 2001, pp. 416–423.

[82] M. Polak, H. Zhang, and M. Pi, "An evaluation metric for image segmentation of multiple objects," *Image and Vision Computing*, vol. 27, no. 8, pp. 1223–1227, Jul. 2009.

[83] A. Hoover, G. Jean-Baptiste, X. Jiang, P. Flynn, H. Bunke, D. Goldgof, K. Bowyer, D. Eggert, A. Fitzgibbon, and R. Fisher, "An experimental comparison of range image segmentation algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 7, pp. 673–689, Jul. 1996.

[84] X. Jiang, C. Marti, C. Irniger, and H. Bunke, "Distance Measures for Image Segmentation Evaluation," *EURASIP Journal on Advances in Signal Processing*, vol. 2006, no. 1, pp. 1–11, 2006.

[85] W. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, 1971.

[86] J. S. Cardoso and L. Corte-Real, "Toward a generic evaluation of image segmentation." *IEEE Transactions on Image Processing*, vol. 14, no. 11, pp. 1773–82, Nov. 2005.

[87] F. Monteiro and A. Campilho, "Distance measures for image segmentation evaluation," *AIP Conference Proceedings*, 2012.

[88] V. Mezaris, I. Kompatsiaris, and M. Strintzis, "Still Image Objective Segmentation Evaluation Using Ground Truth," in *5th COST 276 Workshop*, B. Kovar, J. Prikryl, and M. Vlcek, Eds., 2003, pp. 9–14.

[89] M. Rajab, "Feature extraction of dermatoscopic images by iterative segmentation algorithm," *Journal of X-Ray Science and Technology*, vol. 16, pp. 33–42, 2008.

[90] Y. Zhang and J. Gerbrands, "Segmentation evaluation using ultimate measurement accuracy," *Proceedings of SPIE*, 1992.

[91] R. Cárdenes, R. de Luis-García, and M. Bach-Cuadra, "A multidimensional segmentation evaluation for medical image data." *Computer Methods and Programs in Biomedicine*, vol. 96, no. 2, pp. 108–24, Nov. 2009.

[92] N. R. Pal and D. Bhandari, "Image thresholding: Some new techniques," *Signal Processing*, vol. 33, no. 2, pp. 139–158, Aug. 1993.

[93] C. Hagwood and J. Bernal, "Evaluation of Segmentation Algorithms on Cell Populations Using CDF Curves," *IEEE Transactions on Medical Imaging*, 2011.

[94] Y. Rubner, J. Puzicha, C. Tomasi, and J. M. Buhmann, "Empirical Evaluation of Dissimilarity Measures for Color and Texture," *Computer Vision and Image Understanding*, vol. 84, no. 1, pp. 25–43, Oct. 2001.

[95] D. Olson and D. Delen, *Advanced Data Mining Techniques*, 1st ed. Springer, 2008.

[96] D. Wiley, D. Ghosh, and C. Woodhouse, "Automatic Segmentation of CT scans of Checked Baggage," in *Proceedings of the 2nd International Meeting on Image Formation in X-ray CT*, Salt Lake City, Utah, USA, 2012, pp. 310–313.

[97] P. Southam and R. Harvey, "Texture classification via morphological scale-space: Tex-Mex features," *Journal of Electronic Imaging*, vol. 18, no. 4, p. 043007, Oct. 2009.

[98] R. Unnikrishnan, C. Pantofaru, and M. Hebert, "Toward objective evaluation of image segmentation algorithms." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 929–44, Jun. 2007.

[99] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation." *IEEE Transactions on Medical Imaging*, vol. 23, no. 7, pp. 903–21, Jul. 2004.

[100] H. Benhabiles, J.-P. Vandeborre, G. Lavoue, and M. Daoudi, "A framework for the objective evaluation of segmentation algorithms using a ground-truth of human segmented 3D-models," in *2009 IEEE International Conference on Shape Modeling and Applications*. IEEE, Jun. 2009, pp. 36–43.

[101] "MeVisLab Medical Image Processing and Visualization," Bremen, Germany, 2011.

[102] A. Papoulis, *Probability, Random Variables and Stochastic Processes*, 3rd ed. McGraw-Hill, 1991.

[103] X. Bai, Y. Zhao, Y. Huang, and S. Luo, "Normalized Joint Mutual Information Measure for Image Segmentation Evaluation with Multiple Ground-Truth Images," in *Computer Analysis of Images and Patterns*. Berlin: Springer-Verlag, 2011, pp. 110–117.

[104] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, Mar. 1955.

[105] P. J. Huber, *Robust Statistics*, ser. Wiley Series in Probability and Statistics. Hoboken, NJ, USA: John Wiley & Sons, Inc., Feb. 1981.

[106] T. W. Anderson, "On the Distribution of the Two-Sample Cramer-von Mises Criterion," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1148–1159, Sep. 1962.

[107] S. Kullback and R. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, 1951.

[108] M. Bland, *An Introduction to Medical Statistics*, 3rd ed. Oxford University Press, 2000.